



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학박사학위논문

질량 분석 데이터의 동위 원소
집단에 대한 모델링 및 응용

Modeling and Applications for Isotopic Clusters
in Mass Spectrometric Data

2016 년 2월

서울대학교 대학원

컴퓨터 공학부

윤주영

Abstract

Modeling and Application for Isotopic Clusters in Mass Spectrometric Data

Joo Young Yoon

Department of Computer Science Engineering

The Graduate School

Seoul National University

Mass spectrometry is one of the most robust and powerful analytical tools to identify peptide sequence. It is essential to develop automated methods for analysis of mass spectrometric data since it is impractical to analyze a huge amount of mass spectrometric data manually.

In this thesis, we study high-throughput analysis of mass spectrometry data, especially, determination of isotopic clusters and monoisotopic masses, and peptide quantification using isotope labeling. First, we present a new mathematical model for isotopic distributions of peptides, and propose an algorithm that determines isotopic clusters and monoisotopic masses. Our model uses two types of ratios: intensity ratio of two adjacent peaks and intensity ratio product of three adjacent peaks in an isotopic distribution. We show that those ratios can be approximated as simple functions of a peptide mass, and present an automated algorithm using these functions. We compared the result of our method to the result of well-known THRASH-based implementations. Experimental results show that our method found masses of known peptides than THRASH, especially for peptides whose isotopic distributions deviate significantly from the *averagine* distributions.

Another advantage of our method is the throughput, which is much faster than THRASH that calculates the least-squares fit.

Second, we present a new mathematical model for overlapping isotopic clusters in duplex mTRAQ labeling experiments which is a kind of stable isotope labeling, and propose an algorithm for peptide quantification. It can be easily applied in Trans-Proteomic Pipeline (TPP) instead of XPRESS. For the mTRAQ labeled peptides, it showed more accurate ratios and better standard deviations than XPRESS. Especially, for the peptides that do not contain lysine, the ratio difference between XPRESS and our algorithm became larger as the peptide masses increased.

Finally, we present a new algorithm for peptide quantification in triplex mTRAQ experiments. It is an extension of the previous overlapping model on duplex mTRAQ experiments. We also present an automatic method for determination of the elution areas of peptide. Some peptides have similar atomic masses and elution times, so their elution areas can have overlaps. It is essential to identify the overlap of elution areas and separate them for accurate peptide quantification. We validated the algorithm using standard protein mixture experiments.

Keywords : Mass spectrometry, high-throughput analysis, nomoisotopic mass, isotopic cluster, peptide quantification, mTRAQ labeling

Student Number : 2004-21571

Contents

Abstract	i
1 Introduction	1
1.1 Background	1
1.1.1 Amino acids and isotopes	3
1.1.2 Mass spectrometry experiments	5
1.2 Problem Statement	6
1.3 Previous Works and New Results	7
1.3.1 Determination of isotopic clusters and monoisotopic masses	7
1.3.2 Peptide quantification using stable isotope labeling	9
1.4 Organization	10
2 Determination of Isotopic Clusters and Monoisotopic Masses	11
2.1 Preliminaries	11
2.2 Algorithm	13
2.3 Results	25

3	Peptide Quantification Using mTRAQ Labeling	35
3.1	Preliminaries	35
3.1.1	mTRAQ	35
3.1.2	Tools	36
3.2	Peptide Quantification Using Duplex mTRAQ Labeling	38
3.2.1	Algorithm	38
3.2.2	Results	43
3.3	Peptide Quantification Using Triplex mTRAQ Labeling	58
3.3.1	Algorithm	58
3.3.2	Results	67
4	Conclusion	73
	Bibliography	77

List of Figures

1.1	Structure of an amino acid and condensation of two amino acids to form a peptide bond	4
1.2	Mass spectrometry experiment	4
2.1	Mass spectrum and isotopic cluster	12
2.2	Ratio functions (I_{k+1}/I_k) obtained from stochastic simulation using 100,000 tryptic peptides sampled from Uniprot database	20
2.3	Ratio product functions ($I_k I_{k+2}/I_{k+1}^2$) obtained from stochastic simulation using 100,000 tryptic peptides sampled from Uniprot database	20
2.4	Numbers of clusters of known peptides which were identified by each program Determination of isotopic clusters and monoisotopic masses	26
2.5	Examples where our method determines the correct monoisotopic mass	29
2.6	Examples of overlapping clusters	32
2.7	Execution time of three programs	34
3.1	Overall framework of Trans-Proteomic Pipeline	37
3.2	Examples of overlapping isotopic clusters	39
3.3	Distribution of $\log_{10}(H/L)$ values of peptides from 1:1 human plasma sample	46

3.4	Distribution of $\log_{10}(H/L)$ values of peptides with no lysine	48
3.5	Examples where our method calculated more accurate ratios	50
3.6	Examples of overlapping triplex isotopic clusters	59
3.7	Elution area approximation to normal distribution	62
3.8	Four types of overlaps between chemically different peptides	64
3.9	Manual inspection for the peptides whose computed ratios are different from the expected ratio	68
3.10	Distribution of ratios of peptides for LALBA	70

List of Tables

1.1	Probability of existence of isotopes	3
2.1	Numbers of clusters of 494 known peptides	27
2.2	Result of monoisotopic mass determination for the peptide whose mass is 2296.22 Da	28
3.1	Sample description for duplex mTRAQ.....	44
3.2	Performance comparison between different methods for 1:1 human plasma sample	45
3.3	Expected ratios and computed ratios in S1L1_S2H1 sample	52
3.4	Expected ratios and computed ratios in S1H1_S2L1 sample	53
3.5	Expected ratios and computed ratios in S1L1_S2H5 sample	54
3.6	Expected ratios and computed ratios in S1L5_S2H1 sample	55
3.7	Expected ratios and computed ratios in PLASMA_S1L1_S2H1 sample	56
3.8	Expected ratios and computed ratios in PLASMA_S1H1_S2L1 sample	57
3.9	Standard protein mixtures for triplex mTRAQ	67
3.10	Expected ratios and computed ratios for seven proteins in standard mixtures	72

Chapter 1

Introduction

1.1 Background

Proteomics is the study of proteins, particularly their structures and functions. One of the key technologies for proteomics is peptide sequencing. The primary structure of a protein is a chain of twenty kinds of amino acids. Therefore it can be specified by a sequence of which alphabet size is twenty. Many people usually analyzed peptides which are fragments of proteins since it is difficult to analyze intact proteins.

In early proteomics, Edman degradation is used to identify the sequence of peptides [1]. During the 1990s, with the introduction of soft ionization methods such as electrospray ionization (ESI) [2] and matrix-assisted laser desorption/ionization (MALDI) [3], mass spectrometry (MS) has arisen as one of the most robust and powerful analytical tools to identify peptide sequence. In MS experiments, biomolecules are ionized and their mass is measured by following their specific trajectories in a vacuum system. It innovatively improved proteomic experiments, especially in the throughput.

Since it is impractical to analyze a huge amount of MS data manually, many researchers developed various automated methods for analysis of MS data. There are two types of MS data: mass spectra and tandem mass spectra (also called MS/MS or MS²). By interpreting the mass spectra, we can obtain the molecular masses of peptides, which is fundamental to analyze a corresponding tandem mass spectrum. Many algorithms for interpreting the mass spectra are developed: Mann et al.'s deconvolution algorithm [4], *averagine* [5], Zscore [6], THRASH [7], and so on. And the introduction of tandem MS [8] enables the determination of peptide sequences [9-11].

Another interesting problem is peptide quantification. The quantitative information helps understand the expressional difference of proteins. There are three major experimental strategies for peptide quantification: spectral counting, stable isotope labeling, and label-free quantification [12]. Among these, stable isotope labeling is considered as the most reliable and accurate method. There are various labeling techniques: ICAT [13], SILAC [14], ¹⁸O labeling [15, 16], mTRAQ [17], and so on. Various software tools for isotope labeling have been also developed [18-28].

In this thesis, we study high-throughput analysis of mass spectrometry data, especially, determination of isotopic clusters and monoisotopic masses, and peptide quantification using isotope labeling. In the following subsections, we introduce several terms for mass spectrometry, and general process of mass spectrometry experiments.

mass	¹² C	¹ H	¹⁶ O	¹⁴ N	³² S
+0	98.93	99.989	99.757	99.632	94.93
+1	1.07	0.0115	0.038	0.368	0.76
+2	-	-	0.205	-	4.29
+4	-	-	-	-	0.02

Table 1.1: Probability of existence of isotopes

1.1.1 Amino acids and isotopes

A protein is made of amino acids arranged in a linear chain. Generally, a complete biological molecule is called a protein, and a short amino acid chain is called a peptide. Because there are 20 standard amino acids, we can represent a protein as a sequence (called the primary structure) whose alphabet size is twenty. An amino acid is a molecule which consists of an α -carbon, an amino group, a carboxyl group, and a side chain. Its general formula is $\text{H}_2\text{NCHRCOOH}$, where R is a side chain which has twenty different forms. In Figure 1.1.a, the left part is the amino group and the right part is the carboxyl group. When two amino acids are linked, they form a peptide bond and a molecule of water (Figure 1.1.b).

To sequence a protein, we generally determine the mass of the protein. The mass of a protein is not unique because each element has several different forms called isotopes [29]. For an example, an instance of a protein has four +1 isotopes, its mass is bigger by 4 Da (Dalton, the unit of molecular mass) than an instance of the protein with no isotope. In mass spectrometry, the monoisotopic mass is used as the

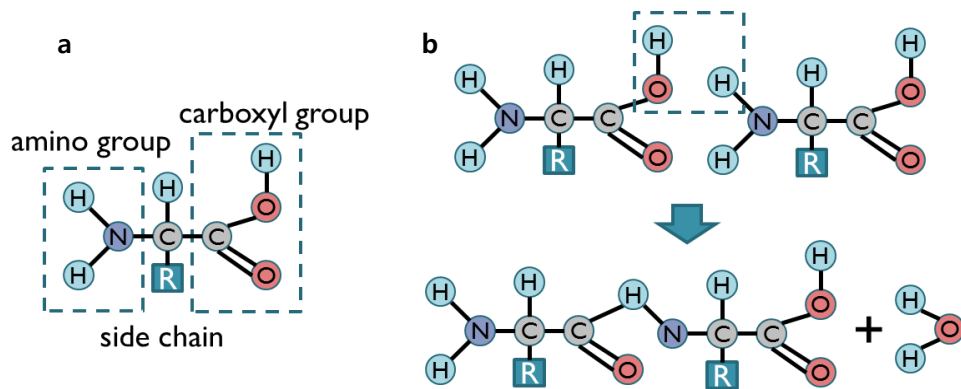


Figure 1.1: (a) Structure of an amino acid and (b) condensation of two amino acids to form a peptide bond

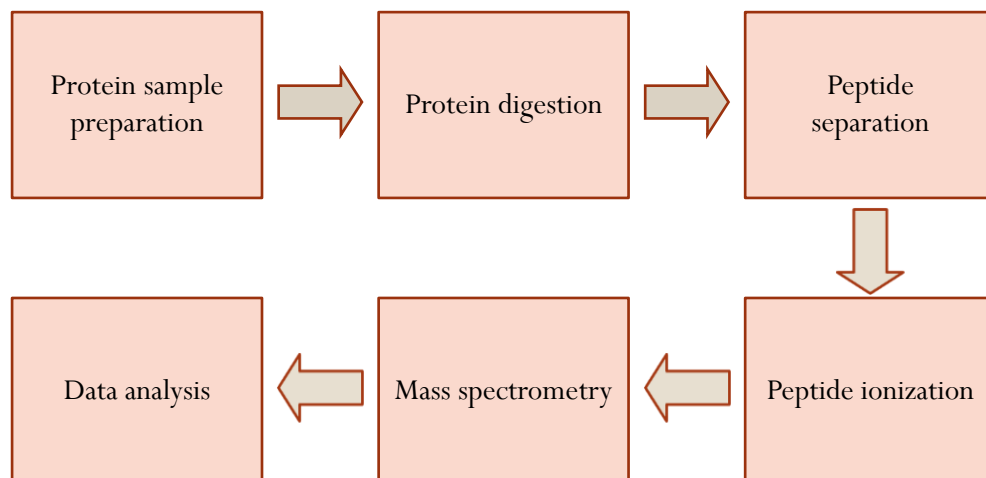


Figure 1.2: Mass spectrometry experiment

mass of a protein or a peptide. It is the sum of the masses of the atoms in a molecule using the mass of the principal isotope for each element. Furthermore, using the probability of existence of isotopes (Table 1.1), we can calculate the isotopic distribution of a molecule [30-32]. It enables us to identify proteins and peptides from complex MS data.

1.1.2 Mass spectrometry experiments

A typical mass spectrometry experiment is processed as shown in Figure 1.2.

First, a sample of interesting proteins is prepared. Because proteins are difficult to handle and might not all be soluble under the same conditions, they are digested by a protease and become short fragments which are called peptides. Most of the mass spectrometers generate the best mass spectrum from peptides that are up to ~20 residues long, rather than from intact proteins. Trypsin is used to convert proteins to peptides in most cases. It predominantly cleaves peptide chains at the carboxyl side of the amino acids lysine or arginine.

The next step is peptide separation. The peptides that are generated by protein digestion are too complex to inject into the mass spectrometer all at once. Therefore, they are injected onto a microscale capillary high-performance liquid chromatography (HPLC) column which is directly connected to the mass spectrometer. The peptides are eluted in order of their hydrophobicity from this column using a solvent gradient of increasing organic content. Hydrophilic peptides might elute immediately and extremely hydrophobic peptides might not elute until most of solvent became organic content.

When a peptide arrives at the end of the column, it is ionized. There are two ionization methods: electrospray ionization (ESI) [2] and matrix-assisted laser

desorption/ionization (MALDI) [3]. In the ESI process, the end of the column is connected to a needle which is held at a high electrical potential (several kV). At the needle tip, the liquid is vaporized and the peptide is subsequently ionized by the action of a strong electric potential. If n protons are coupled to a peptide, its mass increases about $1.0073n$ Da and its charge state becomes n . In the MALDI process, the peptide is mixed with a large amount of matrix molecules. At the end of the column, matrix molecules are sublimed by laser beam and transfer the embedded non-volatile peptide molecules into the gas phase. After numerous ion collisions, singly protonated peptide ions are formed.

Then, peptide ions enter the mass spectrometer. For each unit time, the mass spectrometer determines the mass-to-charge ratios (m/z) of the peptides and generates a mass spectrum which is the recording of the signal intensity of the ion at each m/z value. There are many kinds of mass spectrometers: quadrupole, time of flight (TOF), quadrupole ion trap, Fourier-transform ion cyclotron resonance (FT-ICR), orbitrap, and so on. Using a mass spectrum, the mass spectrometer generates the tandem mass spectra. In tandem mass spectrometry process, a particular peptide ion is isolated. Then energy is imparted to break the peptide and the resulting fragments are generated as a tandem mass spectrum. From the tandem mass spectrum, we finally obtain sequence information of this peptide.

1.2 Problem Statement

In this thesis, we consider the following problems, encountered while analyzing mass spectrometry data.

Determination of isotopic clusters and monoisotopic masses: Given mass spectra each of which is given as an isotopic peak list, find isotopic clusters and determine their monoisotopic masses without peptide sequences.

Peptide quantification using stable isotope labeling: Given a mass spectra each of which is given as an isotopic peak list and a list of peptides which are labeled using two or more type of stable isotope label, determine relative ratios of the given peptides.

1.3 Previous Works and New Results

1.3.1 Determination of isotopic clusters and monoisotopic masses

Determining isotopic clusters and their monoisotopic masses is the first step in interpreting complex mass spectra generated by high-resolution mass spectrometers. Accurate determination of the isotopic clusters increases the quantitative information of the peptides. The monoisotopic masses are used to support the analysis of tandem mass spectra. Furthermore it is possible to improve the selection of isolated peaks in tandem mass spectra if it can make “on-the-fly” determination of the monoisotopic masses.

The first algorithm to determine molecular mass (not monoisotopic mass) was suggested by Mann et al. in 1989 [4]. At that time, there is little information about the feature of mass spectra and isotopic clusters. Because intact proteins are used instead of peptides, each molecule is seen at various charge states. They suggest a new method to find correct charge states.

In 1995, Senko et al. introduced new notion of an average amino acid called “*averagine*” and suggested an algorithm for determining monoisotopic masses using *averagine* [5]. By using the statistical occurrences of the amino acids from the PIR protein database, they obtained an average amino acid of which molecular formula is $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$ and an average molecular mass is 111.1254 Da.

In 1998, Zhang and Marshall proposed new algorithm Zscore [6] for charge state determination and identification of isotopic clusters. There are two notable improvements. First, it is fully automated and does not need user input during the process. Second, it uses new charge determination method which is robust for high charge states.

Horn et al. suggest a fully automated algorithm THRASH [7], which is one of the most widely used algorithms for determining monoisotopic masses today. THRASH is implemented in ICR2LS and Decon2LS program with some modification (<http://ncrr.pnl.gov/software/>). It employs the Fourier transform/Patterson method for charge state determination and least-squares fitting to compare a peak cluster with an *averagine* isotopic distribution.

Many other programs were also developed: ESI-ISOCONV [33], MATCHING [34], PepList [35], LASSO [36], AID-MS [37], and so on.

In this thesis, we present a mathematical model for isotopic distributions of peptides and an effective interpretation algorithm: RAPID [38]. Our model uses two types of ratios: intensity ratio of two adjacent peaks and intensity ratio product of three adjacent peaks in an isotopic distribution. We show that those ratios can be approximated as simple functions of a peptide mass. On the basis of our model, we present an innovative algorithm that determines isotopic clusters and monoisotopic masses.

1.3.2 Peptide quantification using stable isotope labeling

Peptide quantification is one of many interesting computational problems in MS. Stable isotope labeling is considered as the most reliable and accurate method for peptide quantification. There are various labeling techniques: ICAT [13], SILAC [14], ^{18}O labeling [15, 16], mTRAQ [17], and so on. SILAC exploits metabolic labeling of cultured cells in which the medium is supplemented with amino acids containing stable isotopes. ICAT incorporates isotopes on the thiol moiety of cysteine chemically. Similarly, up to 2 ^{18}O atoms can be incorporated into carboxyl groups of peptides by digestion with proteases in the presence of H_2^{18}O . Numerous computational tools for the stable isotope labeling have also been developed, including XPRESS [18], ASAPRatio [19], STEM [20], ZoomQuant [21], MSInspect [22], Multi-Q [23], Q3 [24], VIPER [25], MaxQuant [26], Census [27], and IEMM [28].

One of the major obstacles to accurate peptide quantification using isotope labeling is the overlap of isotopic clusters. There are two types of overlap problems, one is the overlap between differently labeled peptides, and the other is the overlap between chemically different peptides. The former can happen when the mass difference between labels is very small. In mTRAQ experiments, the mass difference between differently labeled peptides is 4 Da if the original peptide has no lysine, so it is important to separate their isotopic clusters correctly. The latter could be found in all kinds of MS-based experiments. For peptide quantitation, most of the times we are interested in relative quantitation of peptides whose amino acid sequences are known. When we know the sequences of peptides of interest, there are better chances to recognize the overlaps from differential labeling by comparing them to the theoretical isotopic distributions.

In this thesis, we present a new algorithm to improve the accuracy of peptide quantification when mTRAQ labeling is used. Most of our analysis is performed using Trans-Proteomic Pipeline (TPP) except that we use our new algorithm instead of XPRESS to quantify the ratios of peptides. We first present a new data analysis algorithm QuadQuant for peptide quantification in duplex mTRAQ experiments [39]. We identify isotopic clusters of labeled peptides and separate them using a mathematical equation modelling of overlapped isotopic cluster. Then, we extend it to triplex mTRAQ experiments [40]. We also designed an automatic determination algorithm for the elution area of peptides, which could recognize the overlap between chemically different peptides.

1.4 Organization

The rest of this thesis is organized as follows. In Chapter 2, we present a new isotopic distribution model and a new algorithm for determining isotopic clusters and monoisotopic masses. In Chapter 3, we present a new algorithm for peptide quantification using mTRAQ labeling by separating overlapping isotopic clusters. Finally, we conclude in Chapter 4.

Chapter 2

Determination of Isotopic Clusters and Monoisotopic Masses

2.1 Preliminaries

Let $A = \{C, H, N, O, S\}$ be the set of atoms that compose a peptide. For each atom $X \in A$, let X_a denote the $+a$ isotope of an atom X , and P_{X_a} denote its existential probability. For example, $P_{C_1} = 0.01107$ because 1.107% of carbon atoms in nature are $+1$ isotopes. $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$ denotes the elemental composition of a peptide where n_X is the number of atom X in the peptide.

Because of the isotopes, the mass of a peptide $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$ is not unique. If an instance of the peptide has four $+1$ isotopes, its mass is bigger by 4 Da than an instance of the peptide with no isotopes. The set of peaks generated by various instances of a peptide is called the isotopic cluster of the peptide (Figure 2.1). In an isotopic cluster, each peak is separated by 1 Da (average value 1.00235 Da [7, 37]).

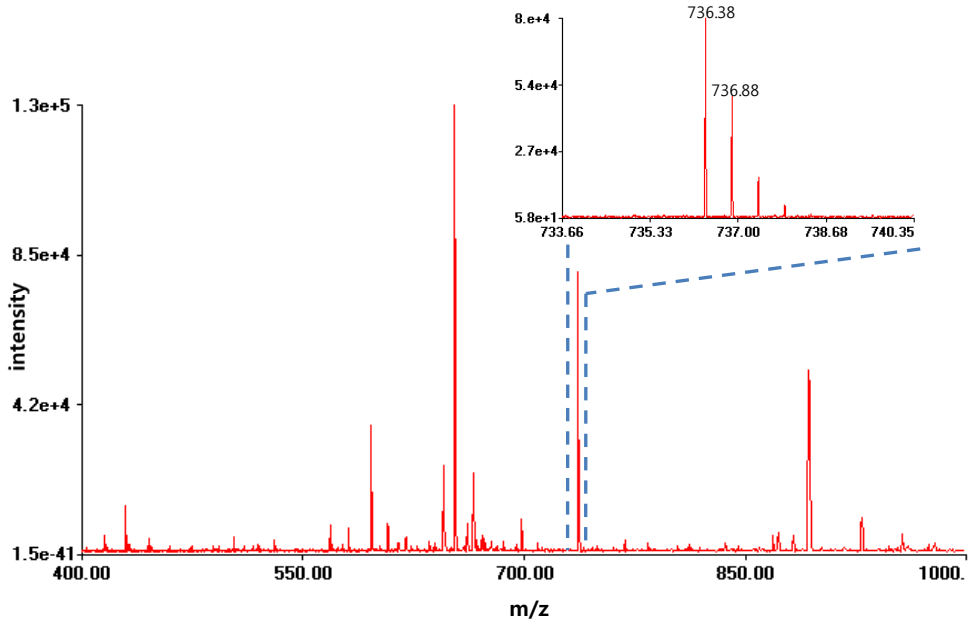


Figure 2.1: Mass spectrum and isotopic cluster

In mass spectrometry experiments, peptides are ionized and injected into enter a mass spectrometer. For each unit time, the mass spectrometer determines the mass-to-charge ratios (m/z) of the peptides and generates a mass spectrum which is the recording of the signal intensity of the ion at each m/z value. The signal intensity at x Th (thomson, the unit of m/z) means the number of ions of which m/z values is x . The mass of a peptide with no isotope is called the monoisotopic mass of the peptide. For an example, a peptide of which monoisotopic mass is 1470.75 Da and charge state is two would be seen at $(1470.75 + 2 \times 1.0073) / 2 = 736.38$ Th in the mass spectrum (see the highlighted peak in Figure 2.1).

2.2 Algorithm

Isotopic Distribution Model

We define an isotopic distribution of a peptide as the theoretical masses and intensities of the peaks generated by all instances of the peptide. Let I_k denote the intensity of the k -th, $k \geq 0$, peak in an isotopic distribution. Specifically, intensity I_0 is the intensity of the monoisotopic peak and I_k , $k \geq 1$, is the intensity of the peak whose mass difference from the monoisotopic peak is k . We model I_k as in Lemma 1, using the existential probability of the peptide instance whose mass is bigger by k Da than the peptide instance with no isotopes.

Lemma 1. *The intensity I_k in an isotopic distribution approximates to*

$$I_k = I_0 \sum_{k_1+2k_2+4k_4=k} \frac{T_1^{k_1} T_2^{k_2} T_4^{k_4}}{k_1! k_2! k_4!},$$

$$\text{where } T_1 = \sum_X \frac{n_X P_{X_1}}{P_{X_0}}, \quad T_2 = \sum_X \frac{n_X P_{X_2}}{P_{X_0}}, \quad \text{and} \quad T_4 = \sum_X \frac{n_X P_{X_4}}{P_{X_0}}.$$

Proof. For a peptide $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$, we can compute I_k by the coefficient of x^k in the expansion of the following polynomial.

$$P(x) = (P_{C_0} + P_{C_1}x)^{n_C} (P_{H_0} + P_{H_1}x)^{n_H} (P_{N_0} + P_{N_1}x)^{n_N} \\ (P_{O_0} + P_{O_1}x + P_{O_2}x^2)^{n_O} (P_{S_0} + P_{S_1}x + P_{S_2}x^2 + P_{S_4}x^4)^{n_S}$$

That is, intensity I_k in an isotopic distribution of a peptide is regarded as the sum of existential probabilities of all peptide instances with mass difference k . Intensity I_0 is the probability of there being no isotopes in $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$, which is the constant term of polynomial $P(x)$, defined as follows.

$$I_0 = P_{C_0}^{n_C} P_{H_0}^{n_H} P_{N_0}^{n_N} P_{O_0}^{n_O} P_{S_0}^{n_S}$$

Intensity I_1 is the probability of there being only one +1 isotope, which is the coefficient of x in $P(x)$, and I_2 is the probability of there being two +1 isotopes or one +2 isotope, which is the coefficient of x^2 , and they are defined as follows.

$$\begin{aligned} I_1 &= n_C P_{C_0}^{n_C-1} P_{C_1} P_{H_0}^{n_H} P_{N_0}^{n_N} P_{O_0}^{n_O} P_{S_0}^{n_S} + n_H P_{C_0}^{n_C} P_{H_0}^{n_H-1} P_{H_1} P_{N_0}^{n_N} P_{O_0}^{n_O} P_{S_0}^{n_S} + \dots \\ &= I_0 \sum_{X \in A} n_X \frac{P_{X_1}}{P_{X_0}} \end{aligned}$$

$$\begin{aligned} I_2 &= \binom{n_C}{2} P_{C_0}^{n_C-2} P_{C_1}^2 P_{H_0}^{n_H} P_{N_0}^{n_N} P_{O_0}^{n_O} P_{S_0}^{n_S} + \binom{n_H}{2} P_{C_0}^{n_C} P_{H_0}^{n_H-2} P_{H_1}^2 P_{N_0}^{n_N} P_{O_0}^{n_O} P_{S_0}^{n_S} \\ &\quad + \dots + \binom{n_S}{2} P_{C_0}^{n_C} P_{H_0}^{n_H} P_{N_0}^{n_N} P_{O_0}^{n_O} P_{S_0}^{n_S-2} P_{S_1}^2 \\ &\quad + n_C n_H P_{C_0}^{n_C-1} P_{C_1} P_{H_0}^{n_H-1} P_{H_1} P_{N_0}^{n_N} P_{O_0}^{n_O} P_{S_0}^{n_S} \\ &\quad + n_C n_N P_{C_0}^{n_C-1} P_{C_1} P_{H_0}^{n_H} P_{N_0}^{n_N-1} P_{N_1} P_{O_0}^{n_O} P_{S_0}^{n_S} \\ &\quad + \dots \\ &\quad + n_O P_{C_0}^{n_C} P_{H_0}^{n_H} P_{N_0}^{n_N} P_{O_0}^{n_O-1} P_{O_2} P_{S_0}^{n_S} + n_S P_{C_0}^{n_C} P_{H_0}^{n_H} P_{N_0}^{n_N} P_{O_0}^{n_O} P_{S_0}^{n_S-1} P_{S_2} \\ &= I_0 \left(\sum_{X \in A} \binom{n_X}{2} \frac{P_{X_1}^2}{P_{X_0}^2} + \sum_{\substack{X, Y \in A \\ X \neq Y}} n_X n_Y \frac{P_{X_1} P_{Y_1}}{P_{X_0} P_{Y_0}} + \sum_{X \in A} n_X \frac{P_{X_2}}{P_{X_0}} \right) \end{aligned}$$

Now consider intensity I_k for an arbitrary $k \geq 1$. The instances with a mass difference $k = k_1 + 2k_2 + 4k_4$ consist of all the instances with k_1 isotopes of +1Da, k_2 isotopes of +2 Da, and k_4 isotopes of +4 Da. For a peptide instance, let t_{X_1} , t_{X_2} , and t_{X_4} be the number of +1, +2, and +4 isotopes of atom X , respectively. Then, the probability of all instances with given k_1 , k_2 , and k_4 is the sum of the probabilities of there being t_{X_1} , t_{X_2} , and t_{X_4} isotopes for each

atom X such that the sum of t_{X_1} for all atoms X is k_1 , that of t_{X_2} is k_2 , and that of t_{X_4} is k_4 . I_k is the probability sum of all combinations of k_1 , k_2 , and k_4 such that $k = k_1 + 2k_2 + 4k_4$ as follows.

$$I_k = \sum_{k_1+2k_2+4k_4=k} \sum_{t_{C_1}+t_{H_1} \dots + t_{S_1}=k_1} \sum_{t_{C_2}+t_{H_2} \dots + t_{S_2}=k_2} \sum_{t_{C_4}+t_{H_4} \dots + t_{S_4}=k_4} \prod_X \binom{n_X}{t_{X_1} \quad t_{X_2} \quad t_{X_4}} P_{X_1}^{t_{X_1}} P_{X_2}^{t_{X_2}} P_{X_4}^{t_{X_4}} P_{X_0}^{n_X - t_{X_1} - t_{X_2} - t_{X_4}}$$

Since n_X (the number of atom X) is much larger than t_{X_1} , t_{X_2} , and t_{X_4} in practice, we employ the following approximation.

$$\binom{n_X}{t_{X_1} \quad t_{X_2} \quad t_{X_4}} = \frac{n_X!}{(n_X - t_{X_1} - t_{X_2} - t_{X_4})! t_{X_1}! t_{X_2}! t_{X_4}!} \approx \frac{n_X^{t_{X_1}+t_{X_2}+t_{X_4}}}{t_{X_1}! t_{X_2}! t_{X_4}!}$$

Then, we obtain the approximation of I_k in Lemma 1 by algebraic manipulations.

$$\begin{aligned} I_k &\approx \sum_{k_1+2k_2+4k_4=k} \sum_{t_{X_1}} \sum_{t_{X_2}} \sum_{t_{X_4}} \prod_X \frac{n_X^{t_{X_1}+t_{X_2}+t_{X_4}}}{t_{X_1}! t_{X_2}! t_{X_4}!} P_{X_1}^{t_{X_1}} P_{X_2}^{t_{X_2}} P_{X_4}^{t_{X_4}} P_{X_0}^{n_X - t_{X_1} - t_{X_2} - t_{X_4}} \\ &= \sum_{k_1+2k_2+4k_4=k} \sum_{t_{X_1}} \sum_{t_{X_2}} \sum_{t_{X_4}} \prod_X P_{X_0}^{n_X} \frac{1}{t_{X_1}!} \left(\frac{n_X P_{X_1}}{P_{X_0}} \right)^{t_{X_1}} \frac{1}{t_{X_2}!} \left(\frac{n_X P_{X_2}}{P_{X_0}} \right)^{t_{X_2}} \frac{1}{t_{X_4}!} \left(\frac{n_X P_{X_4}}{P_{X_0}} \right)^{t_{X_4}} \\ &= \sum_{k_1+2k_2+4k_4=k} \prod_X P_{X_0}^{n_X} \left(\sum_{t_{X_1}} \prod_X \frac{1}{t_{X_1}!} \left(\frac{n_X P_{X_1}}{P_{X_0}} \right)^{t_{X_1}} \right) \left(\sum_{t_{X_2}} \prod_X \frac{1}{t_{X_2}!} \left(\frac{n_X P_{X_2}}{P_{X_0}} \right)^{t_{X_2}} \right) \left(\sum_{t_{X_4}} \prod_X \frac{1}{t_{X_4}!} \left(\frac{n_X P_{X_4}}{P_{X_0}} \right)^{t_{X_4}} \right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{k_1+2k_2+4k_4=k} \frac{\prod_X P_{X_0}^{n_X}}{k_1!k_2!k_4!} \left(\sum_{t_{X_1}} k_1! \prod_X \frac{1}{t_{X_1}!} \left(\frac{n_X P_{X_1}}{P_{X_0}} \right)^{t_{X_1}} \right) \left(\sum_{t_{X_2}} k_2! \prod_X \frac{1}{t_{X_2}!} \left(\frac{n_X P_{X_2}}{P_{X_0}} \right)^{t_{X_2}} \right) \left(\sum_{t_{X_4}} k_4! \prod_X \frac{1}{t_{X_4}!} \left(\frac{n_X P_{X_4}}{P_{X_0}} \right)^{t_{X_4}} \right) \\
&= \sum_{k_1+2k_2+4k_4=k} \frac{I_0}{k_1!k_2!k_4!} \left(\sum_X \frac{n_X P_{X_1}}{P_{X_0}} \right)^{k_1} \left(\sum_X \frac{n_X P_{X_2}}{P_{X_0}} \right)^{k_2} \left(\sum_X \frac{n_X P_{X_4}}{P_{X_0}} \right)^{k_4}. \quad \square
\end{aligned}$$

For example, when $k_1 + 2k_2 + 4k_4 = 4$, there are four cases: four +1 isotopes ($k_1 = 4, k_2 = 0, k_4 = 0$); two +1 isotopes and one +2 isotopes ($k_1 = 2, k_2 = 1, k_4 = 0$); two +2 isotopes ($k_1 = 0, k_2 = 2, k_4 = 0$); and one +4 isotopes ($k_1 = 0, k_2 = 0, k_4 = 1$). Hence I_4 approximates to $I_0 \left(\frac{T_1^4}{4!} + \frac{T_1^2 T_2}{2!} + \frac{T_2^2}{2!} + T_4 \right)$.

Now we want to simplify further the mathematical form of the intensity I_k in Lemma 1. We assume the linearity between mass m and the numbers of atoms, i.e., $n_X \approx a_X m$ where a_X is a constant for each atom X , which may have a range of values according to elemental compositions of peptides. If each n_X is linear in m , then T_1 , T_2 and T_4 are also linear in mass m and I_k becomes a polynomial of m . In the representation of I_k by T_1 , T_2 and T_4 in Lemma 1, the degree of T_1 determines that of I_k , which is k , because the term with highest degree is $\frac{T_1^k}{k!}$

from the case of k isotopes of +1 Da.

Lemma 2. *In an isotopic distribution of a peptide $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$, intensity I_k approximates to a polynomial of mass m with degree k , i.e., $I_k = c_k m^k + c_{k-1} m^{k-1} + \dots + c_1 m + c_0$.*

Due to variations in elemental compositions, each of T_1 , T_2 and T_4 has a range of constants in its linear form. For example, consider the extreme case that a peptide consists of one kind of amino acid: peptides of phenylalanine (F: C_9H_9NO) give the maximum $T_1 = 6.97 \times 10^{-4}m$ and peptides of aspartic acid (D: $C_4H_5NO_3$) the minimum $T_1 = 4.23 \times 10^{-4}m$. The average $T_1 = 5.43 \times 10^{-4}m$ is computed from the *averagine* $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$. Note that the *averagine* model fixes T_1 , T_2 and T_4 as the average values for all values of m . However, we obtain both minimum and maximum of T_1 , T_2 and T_4 as linear forms in addition to their averages. From the ranges of values T_1 , T_2 and T_4 can take, we can estimate the range of I_k .

Ratio Functions and Ratio Product Functions

Based on the approximation of I_k given above, we first show that an intensity ratio, I_{k+1}/I_k , can be approximated to a linear function of peptide mass and that an intensity ratio product, $I_k I_{k+2}/I_k^2$, to a constant function. Recently, a similar model using the intensity ratio was proposed independently, in which I_{k+1}/I_k is modeled by a polynomial of mass [41]. They approximated the intensity ratios as functions of the monoisotopic mass and the number of sulfur atoms which are more complex than ours. We show here that a simple linear approximation of I_{k+1}/I_k suffices for identification of isotopic clusters.

Second, we compute their average, minimum and maximum functions using simulation spectra of tryptic peptides generated from a protein database. The algebraic estimation of min/max functions from T_1 , T_2 and T_4 becomes harder for higher degree k , so we compute them using stochastic simulation. These intensity ratio and ratio product functions are simpler than the intensity itself and reveal more features of isotopic distributions.

From Lemma 2, I_{k+1}/I_k is a ratio of two polynomials of degree $k + 1$ and k . For a sufficiently large mass m , the highest degree terms ($c_{k+1}m^{k+1}$ in I_{k+1} and $c_k m^k$ in I_k) dominate and thus I_{k+1}/I_k approximates to some linear function, $cm + b$.

Theorem 1. *In an isotopic distribution of a peptide $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$, the ratio of two adjacent peaks, I_{k+1}/I_k , can be approximated by a linear function of the peptide mass.*

To determine the constants of the ratio function, $I_{k+1}/I_k = cm + b$, we sampled about 100,000 tryptic peptides of 400Da to 5,200Da generated from UniProt database 8.0 [42] and computed ratio I_{k+1}/I_k for each peptide. Figure 2.2 shows our ratio functions I_{k+1}/I_k for $0 \leq k \leq 3$. For a sufficiently large mass $m \geq 1800$, it can be clearly seen that the intensity ratios can be approximated by linear functions of mass, represented as the solid lines in Figure 2.2, which is in accordance with our theoretical analysis. The solid line, named $R_{\text{avg}}(k, m)$, is computed by linear regression using least-squares fitting in *gnuplot* program (<http://www.gnuplot.info>). The dotted line, $R_{\text{max}}(k, m)$, is the upper bound and the dashed line, $R_{\text{min}}(k, m)$, is the lower bound of the ratios in the graph, also computed by linear regression using least-squares fitting. Note that the min/max functions, $R_{\text{min}}(k, m)$ and $R_{\text{max}}(k, m)$, represent the variation of I_{k+1}/I_k due to elemental composition of peptides of mass m . The average function $R_{\text{avg}}(k, m)$ is very close to the line estimated by *averagine*.

For a small mass $m < 1800$, we use the linear-like quotient of two polynomials with degrees $k + 1$ and k in Lemma 2. Especially, I_1/I_0 has a strong linearity

for all m , because the quotient of I_1/I_0 is cm . The reason for choosing the threshold 1800 is that a peptide within 1800 Da has the first and most abundant peak as its monoisotopic peak. In other words, I_0 is the most abundant and I_{k+1}/I_k , $k \geq 1$, becomes insignificant in the range of $m < 1800$. Note that the model by Valkenborg et al. [41] proposes a refined model of isotopic distributions for low mass peptides by considering the number of sulfurs in the peptides, which explains the tails of ratios in the low mass range. However, our simple model performed well in the experimental data, and we expect that the experimental error in peaks dominates the theoretical error in our model.

In a similar way to Theorem 1, we obtain a constant approximation of the ratio product of three adjacent peaks (i.e., $I_{k+1}/I_k \cdot I_{k+2}/I_{k+1}$). From Lemma 2, degrees of $I_k \cdot I_{k+2}$ and I_{k+1}^2 are the same as $2k + 2$. Hence, $I_k I_{k+2}/I_{k+1}^2$, can be approximated as a constant for peptides of sufficiently large masses.

Theorem 2. *In an isotopic distribution of a peptide $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$, the ratio product of three adjacent peaks, $I_k I_{k+2}/I_{k+1}^2$, can be approximated to a constant.*

Similarly to the ratio functions, we define ratio product functions $RP_{\max}(k, m)$, $RP_{\min}(k, m)$, and $RP_{\text{avg}}(k, m)$, respectively corresponding to the maximum, the minimum and the average values of $I_k I_{k+2}/I_{k+1}^2$. These functions are also computed from the peptide database. We also divide the mass range by 1800 Da and compute the ratio products for two intervals (Figure 2.3).

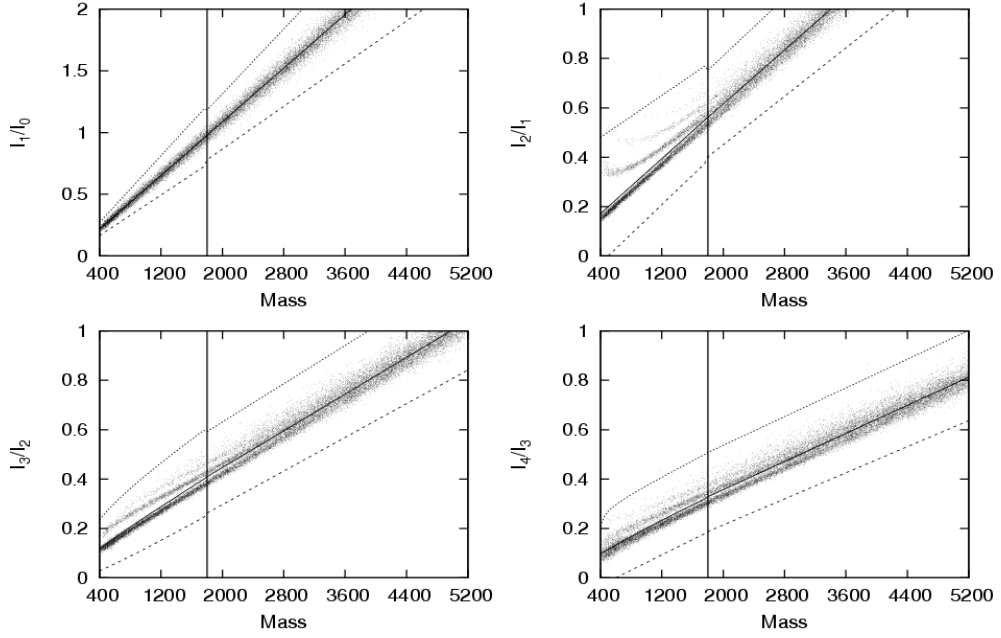


Figure 2.2: Ratio functions (I_{k+1}/I_k) obtained from stochastic simulation using 100,000 tryptic peptides sampled from Uniprot database

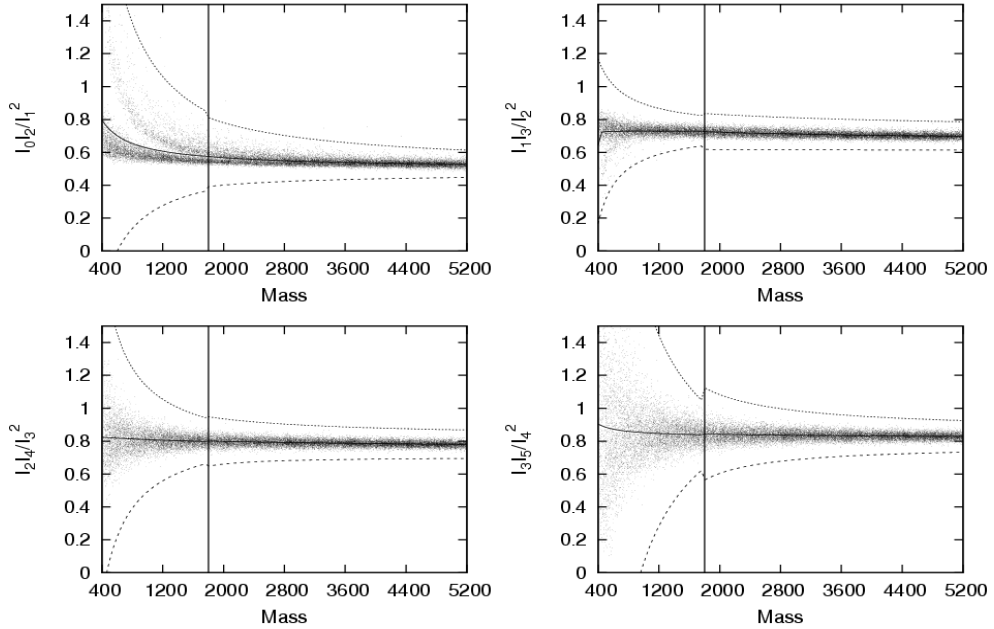


Figure 2.3: Ratio product functions ($I_k I_{k+2}/I_{k+1}^2$) obtained from stochastic simulation using 100,000 tryptic peptides sampled from Uniprot database

Algorithm Overview

We present an algorithm for determining isotopic clusters and their monoisotopic masses from a raw spectrum. Before describing our algorithm, we introduce several cluster names. A *peak cluster* indicates a list of peaks selected from a raw spectrum and sorted in increasing order of m/z . A *pseudo (isotopic) cluster* with charge stage C is a peak cluster such that the m/z difference of every adjacent peak pair in the peak cluster is $1/C$. An *isotopic cluster* with charge state C is a pseudo cluster with charge state C such that the intensity pattern of the pseudo cluster corresponds to that of an isotopic distribution. Our determination algorithm consists of the following four steps: (1) peak picking, (2) pseudo cluster identification, (3) isotopic cluster identification and monoisotopic mass determination, and (4) duplicate cluster removal. We describe the steps one by one.

Peak Picking. We remove noise and select relatively high intensity peaks from the raw spectrum. It should be noted that this step is not closely related to the essence of our algorithm. On the contrary, it is more related to the noise pattern of a mass spectrometer. Thus, any peak picking algorithm that removes well the noise from the raw spectrum can be used. In our experiment, we used the peak picking algorithm of Decon2LS.

Pseudo Cluster Identification. We identify pseudo clusters by scanning the selected peaks from low m/z to high m/z . Every time we examine a peak, we find all the pseudo clusters starting at the peak, in a way that we first find pseudo clusters with a charge state $1+$ and find the other pseudo clusters with higher charge states by incrementing the charge state. We describe how to enumerate all pseudo clusters starting at a peak P with a charge state C . We first enumerate

pseudo clusters with 2 peaks and then pseudo clusters with more peaks: Let X denote the m/z of P : We first find next peaks of P , i.e., peaks in the mass range $[X+(D-E)/C .. X+(D+E)/C]$ where D is the estimated mass difference between two adjacent peaks in an isotopic cluster and E is the error bound. In our experiment, D is 1.00235, which is the mass difference of two adjacent *average* peaks and $E = 10^{-5} * X$, which corresponds to 10 ppm mass accuracy. By pairing P and each next peak of P , we generate all pseudo clusters with 2 peaks. Once pseudo clusters with 2 peaks are enumerated, we enumerate pseudo clusters with 3 peaks by extending the pseudo clusters with 2 peaks to the second next peaks of P . In this way, we can enumerate all pseudo clusters starting at a peak P with a charge state C .

Isotopic Cluster Identification and Monoisotopic Mass Determination. From the pseudo clusters, we identify isotopic clusters, whose intensity patterns are similar to those of isotopic distributions. For each pseudo cluster, we determine whether it is an isotopic cluster or not by checking the intensity ratio of every adjacent peaks and the intensity ratio product of every three adjacent peaks in the pseudo cluster. In determining isotopic clusters, we also consider the case that some peaks are missing in pseudo clusters because sometimes the monoisotopic and its neighboring peaks are as small in their intensities as the noise level and they may be missing from a pseudo cluster. Our algorithm allows up to three leftmost peaks to be missing in a pseudo cluster. More specifically, we calculate scores for four cases (in which we assume that we miss zero to three leftmost peaks) and select the case with the highest score. If the score of the selected pseudo cluster is above zero, it means that the most of ratios and ratio products range from $R_{\max}(k, m)$ to $R_{\min}(k, m)$ and from $RP_{\max}(k, m)$ to $RP_{\min}(k, m)$, respectively.

Therefore, we select pseudo clusters whose scores are larger than zero as isotopic clusters, and other pseudo clusters are discarded.

Score calculation for a pseudo cluster starts with monoisotopic mass calculation. The monoisotopic mass, denoted by m , is computed from the most abundant peak in the pseudo cluster. If the most abundant peak is the q -th peak in the pseudo cluster and p peaks are assumed to be missing, m is computed as follows.

$$m = \text{mass of the } q\text{-th peak} - 1.00235 * (q+p-1)$$

The score of a pseudo cluster with p peaks assumed missing is as follows.

$$\text{Score} = \sum_{k=0}^{n-2} \text{scoreR}(k, p, m) + \sum_{k=0}^{n-3} \text{scoreRP}(k, p, m), 0 \leq p \leq 3$$

where n is the number of peaks in the pseudo cluster.

The score is the sum of ratio score $\text{scoreR}(k, p, m)$ defined on every two adjacent peaks and ratio product score $\text{scoreRP}(k, p, m)$ defined on every three adjacent peaks in a pseudo cluster. Let intensity I'_k be the intensity of the $(k+1)$ -st peak in a pseudo cluster (Note that I'_k corresponds to I_{k+p} in the isotopic distribution). The ratio score $\text{scoreR}(k, p, m)$ measures the similarity of the intensity ratio I'_{k+1}/I'_k to the intensity ratio I_{k+p+1}/I_{k+p} in the isotopic distribution whose monoisotopic mass is m :

$$\text{scoreR}(k, p, m) = \begin{cases} 1 - \frac{I'_{k+1}/I'_k - R_{\text{avg}}(k+p, m)}{R_{\text{max}}(k+p, m) - R_{\text{avg}}(k+p, m)} & \text{if } \frac{I'_{k+1}}{I'_k} > R_{\text{avg}}(k+p, m), \\ 1 - \frac{R_{\text{avg}}(k+p, m) - I'_{k+1}/I'_k}{R_{\text{avg}}(k+p, m) - R_{\text{min}}(k+p, m)} & \text{otherwise.} \end{cases}$$

The ratio score function consists of two linear function fragments of the ratio I'_{k+1}/I'_k that is designed to have the maximum value 1 when the ratio is

$R_{\text{avg}}(k + p, m)$, and to have 0 when the ratio is $R_{\text{max}}(k + p, m)$ or $R_{\text{min}}(k + p, m)$. In addition, the score has negative values when the ratio is higher than $R_{\text{max}}(k + p, m)$ or lower than $R_{\text{min}}(k + p, m)$.

The ratio product score $\text{scoreRP}(k, p, m)$ measures the similarity of the intensity ratio product $I'_k I'_{k+2} / I'_{k+1}{}^2$ to the intensity ratio product $I_{k+p} I_{k+p+2} / I_{k+p+1}{}^2$ in an isotopic distribution whose monoisotopic mass is m :

$$\text{scoreRP}(k, p, m) = \begin{cases} 1 - \frac{I'_k I'_{k+2} / I'_{k+1}{}^2 - RP_{\text{avg}}(k + p, m)}{RP_{\text{max}}(k + p, m) - RP_{\text{avg}}(k + p, m)} & \text{if } \frac{I'_k I'_{k+2}}{I'_{k+1}{}^2} > RP_{\text{avg}}(k + p, m), \\ 1 - \frac{RP_{\text{avg}}(k + p, m) - I'_k I'_{k+2} / I'_{k+1}{}^2}{RP_{\text{avg}}(k + p, m) - RP_{\text{min}}(k + p, m)} & \text{otherwise.} \end{cases}$$

Duplicate Cluster Removal. Because we consider all possible pseudo clusters, many pseudo clusters can be generated from a single isotopic cluster. Suppose that there are five peaks and adjacent peaks are separated by 0.5 Th. In this case, a pseudo cluster consisting of five peaks (with charge state 2+), a pseudo cluster consisting of four peaks (missing the first peak) and a pseudo cluster consisting of three peaks (with charge state 1+) can be generated. We call these clusters “duplicate clusters” and select one of them. (They are not overlapping clusters.) Generally, if two clusters shares one or more peaks and the charge state of one is a multiple of another’s, they are duplicate clusters. Then we remove one of them as follows. First, we remove an isotopic cluster whose most abundant peak is smaller than another’s. If the most abundant peaks are the same, an isotopic cluster with the lower charge state is removed. If their charge states are also the same, the cluster with the lower score is removed.

2.3 Results

We tested our algorithm on a data set from tryptic digests of an 18 protein mixture, “ISB standard protein mix” [43]. To evaluate the performance of our method, we compared it with ICR2LS and Decon2LS, both developed by Smith group at Pacific Northwest National Laboratory (<http://ncrr.pnl.gov/software/>). ICR2LS is a powerful FTICR mass analysis software package. For deisotoping, it basically adapts THRASH. Decon2LS also adapts THRASH, but its algorithm has been modified to increase deisotoping speed while the details of the improvements were not disclosed. All programs were executed on the same PC (Pentium M processor 1.70GHz, 1-GB RAM, Windows XP OS). To be as fair as possible to each program, parameters were set so that each method works on a similar number of total clusters. Our method and Decon2LS use the same peak picking method. The result of each peak picking program contained about twenty five thousand isotopic clusters.

Identification of Known Peptides

In comparing three programs, we counted the number of identified isotope clusters of known peptides whose amino acid sequences were identified by MS/MS. It is difficult, however, to pick out the isotopic clusters of known peptides because the MS data from an LC/MS/MS can contain many peptides whose monoisotopic masses are very similar. Therefore we use the following method to classify peptides. For each known (confidently identified by MS/MS spectrum) peptide, we find isotopic clusters of this peptide at the MS scan where this peptide was identified by tandem MS. If an isotopic cluster has the monoisotopic mass within a

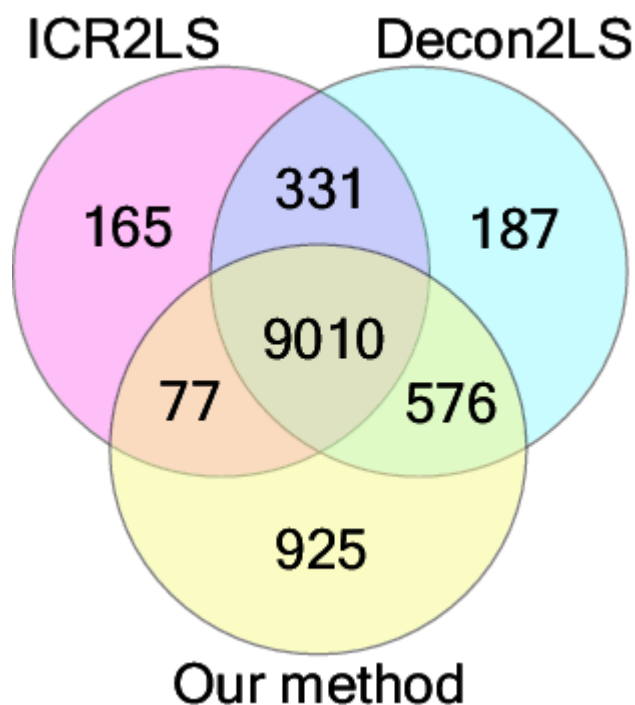


Figure 2.4: Numbers of clusters of known peptides which were identified by each program

mass tolerance of 10 ppm, we consider it a potentially correct isotopic cluster. We also look for this peptide in adjacent scans. If no isotopic cluster is found within any of 10 consecutive scans, the cluster is discarded. We regard these isotopic clusters as true positives.

We counted the isotopic clusters of 494 known peptides. Figure 2.4 shows the number of isotopic clusters identified by each program. It shows a 10.6% improvement over ICR2LS and a 4.8% improvement over Decon2LS. To observe the performance according to the mass, we divided the 494 peptides into 500 Da intervals and counted the number of identified clusters of peptides that belong to each interval (Table 2.1). Our method works well regardless of peptide masses.

Mass	number of peptides	number of clusters		
		Our method	Decon2LS	ICR2LS
~1000	47	790	767	777
~1500	158	2630	2559	2575
~2000	109	2136	2024	1961
~2500	72	1555	1447	1393
~3000	52	1162	1151	1060
~3500	26	963	880	802
~4000	19	969	856	687
~4500	2	42	41	37
~5000	2	30	31	30
5000~	7	311	348	255
Sum	494	10588	10104	9577

Table 2.1: Numbers of clusters of 494 known peptides

There can be various reasons that each program gives different search results. Some clusters are inherently ambiguous and each program can make different judgments. Sometimes the charge states of clusters are determined incorrectly. For all three programs, primary errors are 1-2 Da errors. In THRASH based algorithms, 1-2 Da errors often happen when the position of the most abundant peak of an identified cluster is different from that of *averagine*. On the contrary, our method has low dependency on the most abundant peak. Sometimes THRASH based algorithms determine the monoisotopic mass of an identified isotopic cluster 1 Da larger than the correct mass, even though there exists the correct monoisotopic peak in the spectrum. Such an error is uncommon in our method because adding the monoisotopic peak to the pseudo cluster usually increases the

	Our method	Decon2LS	ICR2LS
2296.22 Da (Correct)	35	27	21
2295.22 Da (-1 Da)	2	1	0
2297.22 Da (+1 Da)	6	10	9
2298.22 Da (+2 Da)	0	1	2
765.40 Da (Wrong CS)	0	2	6
Not found	0	2	5

Table 2.2: Result of monoisotopic mass determination
for the peptide whose mass is 2296.22 Da

score. However, our method also cannot correctly identify several ambiguous cases because it is still based on the shape of the clusters.

Detection of false positives can only be performed by manual inspection because many unidentified peptides are crowded in the spectrum and it is possible that there exists a peptide whose monoisotopic mass is 1 Da different from a known peptide. Here we present several examples in which monoisotopic masses determined by our method are different from masses of other programs. A peptide whose chemical formula is $C_{101}H_{165}N_{29}O_{32}$ and monoisotopic mass is 2296.22 Da is observed in relatively long duration in elution time (from scan no. 3464 to 3565) during LC/MS/MS experiment of the ISB standard peptide mix. The results of mass determination are summarized in Table 2.2. We show four examples in Figure 2.5 where our method determines the correct monoisotopic mass. Circles, diamonds and stars represent the theoretical isotopic distributions of this peptide calculated by each of our method, Decon2LS and ICR2LS, respectively. In Figure 2.5.a, Decon2LS determined the mass of the cluster as 1 Da smaller than the correct theoretical mass because the first peak of the cluster is much larger than the

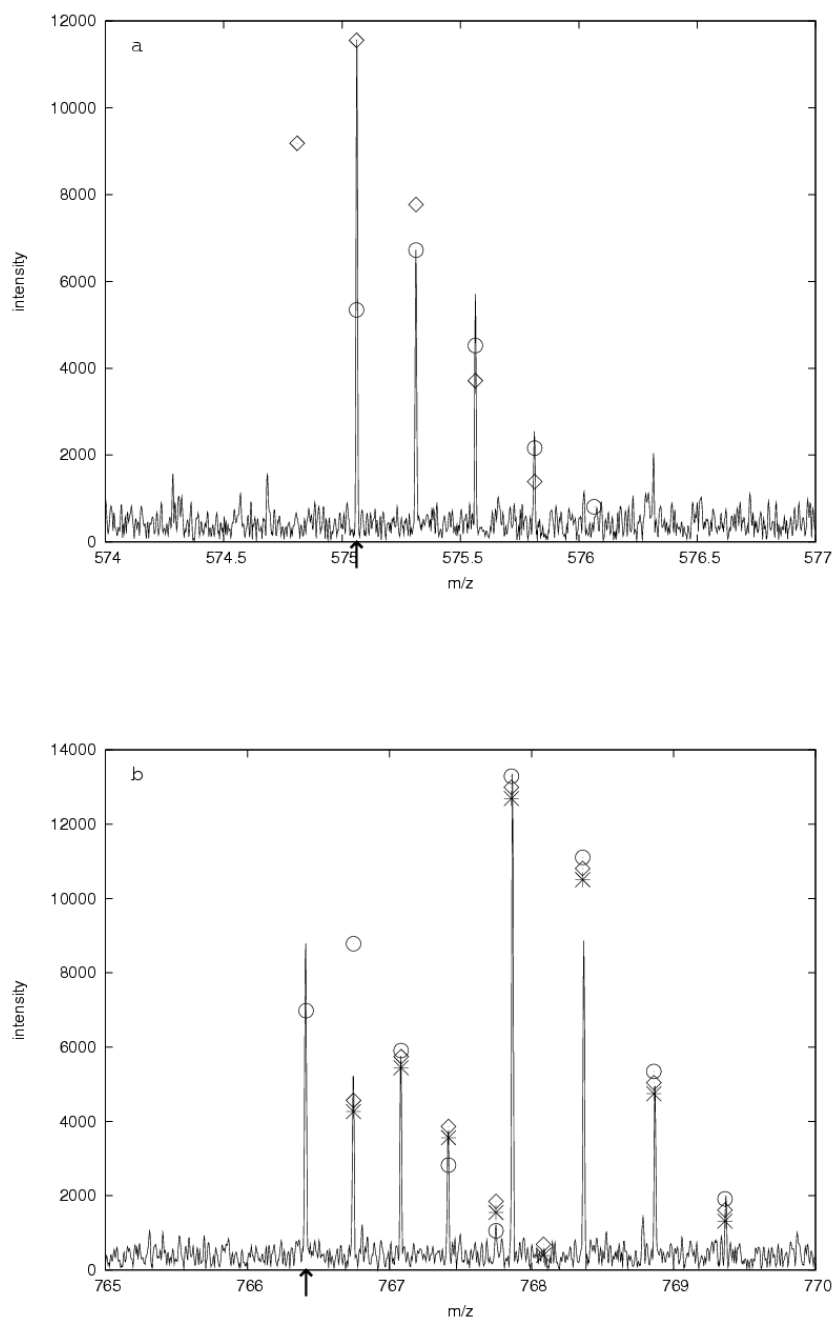


Figure 2.5: Examples where our method determines the correct monoisotopic mass

averagine isotopic distribution. ICR2LS found no cluster in this region. On the other hand, Decon2LS and ICR2LS assigned 2297.22 Da, which is 1 Da larger than the theoretical mass in Figure 2.5.b. Figure 2.5.c is a case where the intensities are close to the noise level. Because the fourth peak appears abnormally large, ICR2LS assigned 2298.22 Da, which is 2 Da larger than the theoretical mass. These examples (Figure 2.5.a~c) show that THRASH algorithm often assigns incorrect mass when the most abundant peak of the identified cluster shows a discrepancy from the *averagine* isotopic distribution. Figure 2.5.d is a case where ICR2LS assigned an incorrect charge state and assigned 765.40 Da as the monoisotopic mass. The clusters that were not found by a program may be found if the parameters are set differently (lowering minimum S/N ratios, for example). However, the different parameter set may well cause false positive determination of other clusters and there is always compromise between the accuracy and computational costs. The highly accurate determination of monoisotopic masses by our method should increase the accuracy peptide identification and decrease false positive peptide identification by MS-based proteomics.

Identification of Overlapping Clusters. Although FT-ICR MS has a high resolving power, there are many overlapping clusters because hundreds of isotopic clusters crowded into a narrow range. Even in these cases it is easy to identify all overlapping isotopic clusters if there is no shared peak. All programs correctly found two isotopic clusters in Figure 2.6.a. However, it is very hard to identify all clusters if isotopic clusters share one or more peaks. THRASH fails to identify all clusters that share one or more peaks, because it deletes the peaks of a cluster when the cluster is determined. Our method can identify overlapping clusters that share one or more peaks in many cases because we consider all possible pseudo clusters

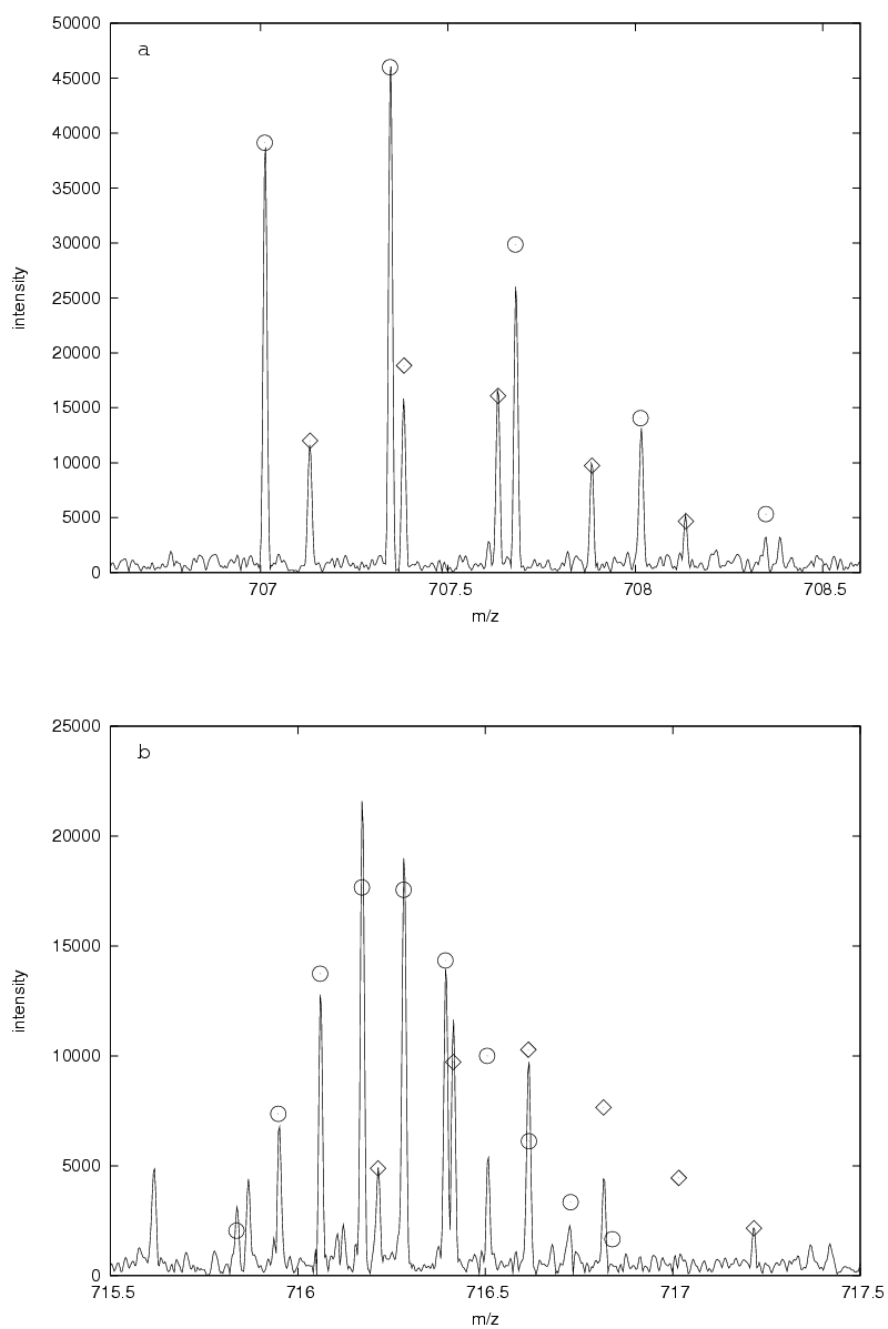


Figure 2.6: Examples of overlapping clusters

and do not delete the peaks of identified clusters. In Figure 2.6.b, the cluster whose monoisotopic mass is 6433.46 Da (diamonds) was identified by all three programs, but the cluster whose monoisotopic mass is 3576.03 Da (stars) was identified only by our method. Both clusters belong to the clusters of 494 known peptides. However, Decon2LS and ICR2LS have failed to identify both because the peak of 716.62 Th is shared by both clusters. Elimination of the 716.62 Th peak results in low match (i.e. low fit number) between the theoretical *average* distribution and the experimental distribution, leading to loss of the mass information.

Execution Time. Another noticeable advantage of our method is its speed. Since our method uses simple ratio functions and ratio product functions that are precomputed, our method can calculate the scores of isotopic clusters much faster than THRASH calculating the least-squares fit on the fly. Execution time for our data set is shown in Figure 2.7. ICR2LS is much slower than other programs. Execution time of our method was similar to that of Decon2LS in deisotoping the first segment data because of the dominant effect of I/O time. We can see a remarkable difference in execution time in analyzing segment 4 data, (almost 5 times faster than Decon2LS,) for which it took the longest time. It must also be noted that the number of peaks obtained by the peak picking step is a major factor in execution time.

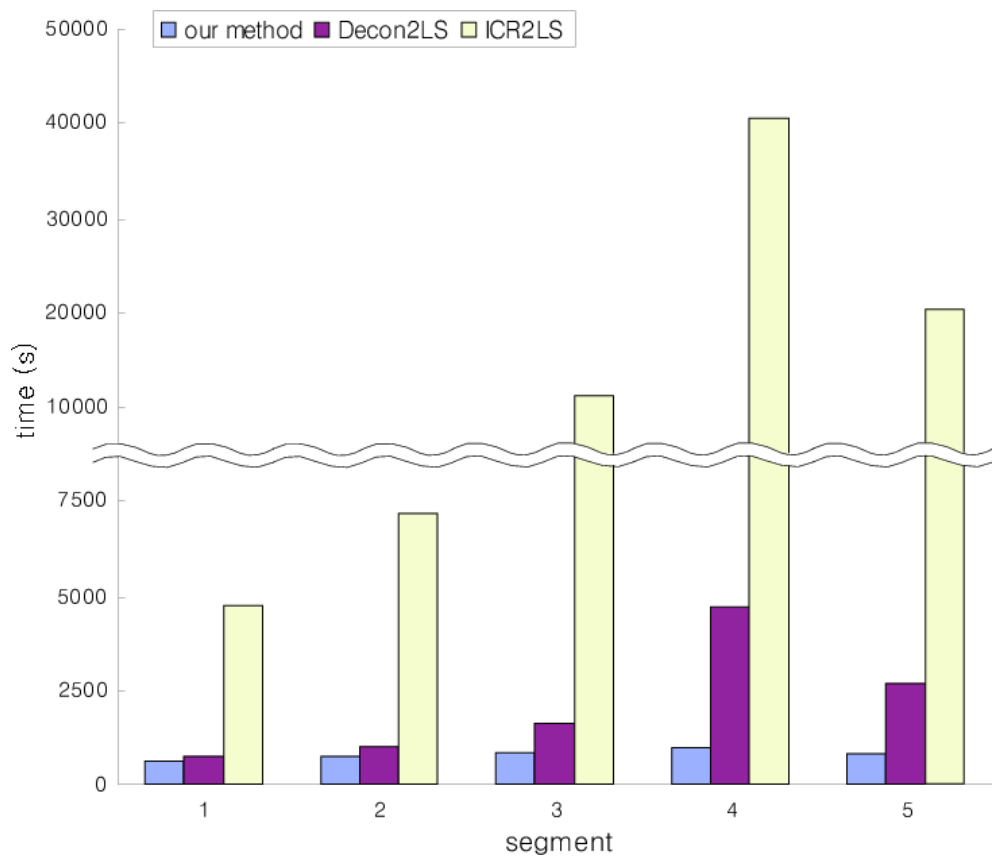


Figure 2.7: Execution time of three programs

Chapter 3

Peptide Quantification Using mTRAQ Labeling

3.1 Preliminaries

3.1.1 mTRAQ

mTRAQ is a non-isobaric variant of the iTRAQ and was originally designed for multiple reaction monitoring (MRM) [17]. The mTRAQ labels were first designed in two chemically identical versions. The heavy-label is identical to the iTRAQ 117 label and its mass is 145 Da. The light-label is chemically identical to the heavy-label, but it has no ^{13}C or ^{15}N , so its mass is 141 Da. They are labeled at lysine residue and N-terminal, therefore, the mass difference between light- and heavy-labeled peptides is $4 + 4 N_L$ Da, where N_L is the number of lysine residue in the peptide. Recently, the mTRAQ has become available in triplex format, where the label with 149 Da is added.

3.1.2 Tools

Our algorithm is designed to be executed in the Trans-Proteomic Pipeline (TPP, <http://tools.proteomecenter.org/software.php>), which is an open source proteomics analysis tool. The overall framework we used is shown in Figure 3.1. For each LC/MS experiment, TPP generates a pepXML file which contains a list of peptides with sequences, tandem scans, charges, and modifications.

SEQUEST [10] is one of the most widely used algorithms to determine the peptide sequence. The main feature of SEQUEST algorithm is a signal processing technique called cross-correlation which measures similarity of two waveforms. SEQUEST retrieves candidate sequences from the database. For all peptides in the database, it calculates the monoisotopic masses of the peptides. A peptide is included to the candidate set if the mass of the peptide falls within a specified mass tolerance. Then, SEQUEST generates the theoretical spectra for all peptides in the candidate set and compares these theoretical spectra to the observed tandem mass spectrum using cross-correlation.

PeptideProphet [44] is a robust and accurate statistical algorithm for validation of peptide identifications made by tandem mass spectrometry and database searching. By employing database search scores, number of tryptic termini, number of missed cleavages, and other information, it assigns a probability of being correct for assigned peptides.

XPRESS [18] is one of algorithms which calculates the relative abundance of labeled peptides, and is included in the current TPP distribution. It reconstructs the light and heavy elution profiles of the precursor ions and determines the elution areas of each peak, and summarizes intensities of monoisotopic peaks in the elution areas.

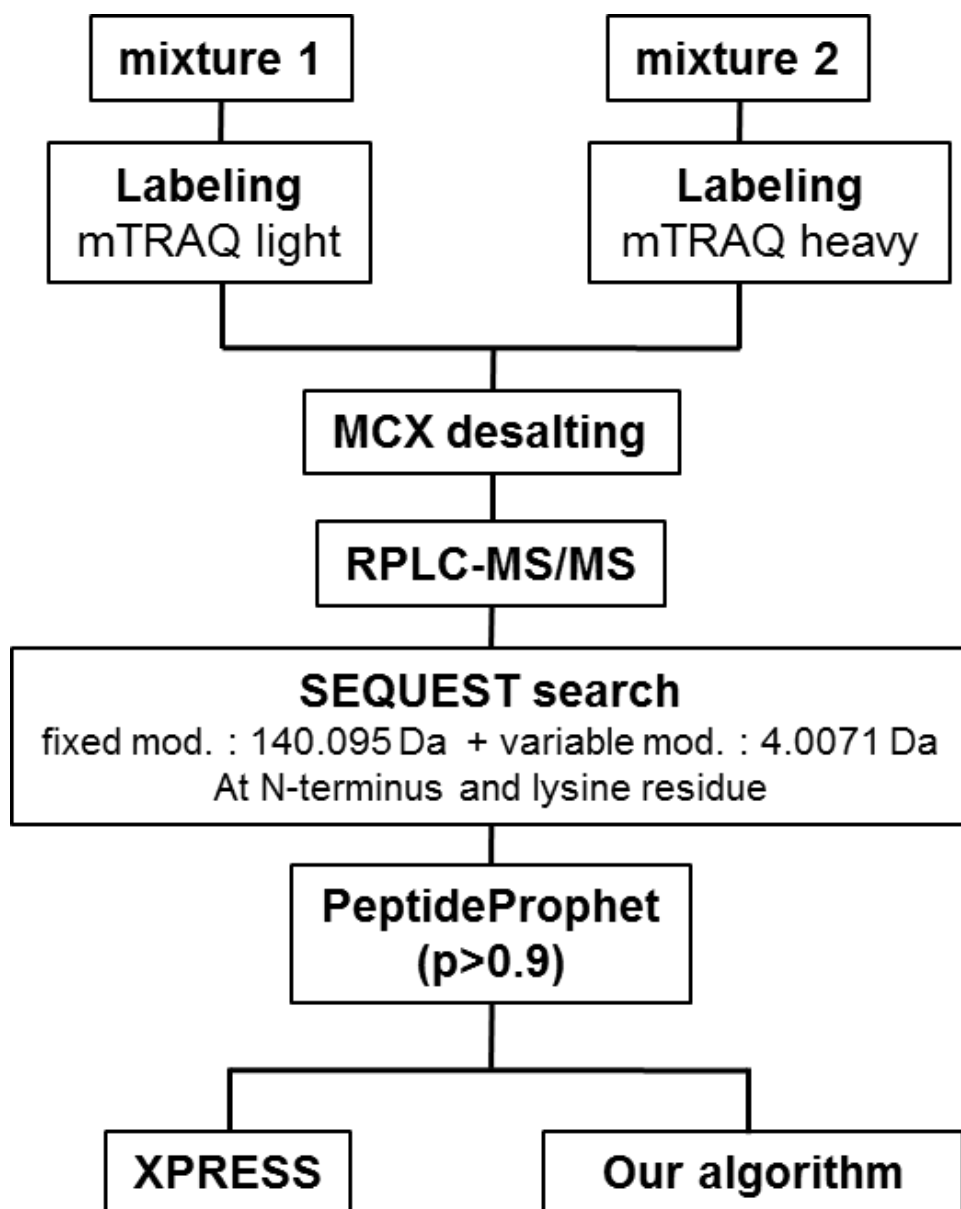


Figure 3.1: Overall framework of Trans-Proteomic Pipeline

It allows the specification of which residues are labeled (such as cysteines for mTRAQ) and what the mass difference of the two isotope labels are (such as 4 Da for mTRAQ).

3.2 Peptide Quantification Using Duplex mTRAQ

Labeling

3.2.1 Algorithm

Model for overlapping isotopic clusters

The mass difference between light and heavy mTRAQ-labeled peptides is about 4 Da if the original peptide has no lysine. These pairs have an overlap in their isotopic clusters if the light-labeled isotopic cluster has five or more isotopic peaks (Figure 3.2). In this thesis, we assume that an isotopic cluster of a peptide has 8 or less peaks. It is reasonable for peptides whose masses are less than 4000 Da because the relative intensity of the ninth peak in the theoretical distribution of an *averagine* [5] whose mass is 4000 Da is only 0.56%. The intensity, I_k , of the k th peak of a theoretical distribution of overlapping isotopic clusters is given as follows:

$$I_k = \begin{cases} L_k & \text{if } k \leq 4, \\ L_k + H_{k-4} & \text{if } 4 < k \leq n, \\ H_{k-4} & \text{if } k > n. \end{cases}$$

where n is the number of peaks in the isotopic distribution of a peptide, L_k is the intensity of the k -th peak of the isotopic distribution of the light-labeled peptide, and H_k is the intensity of the k -th peak of the isotopic distribution of the heavy-labeled peptide.

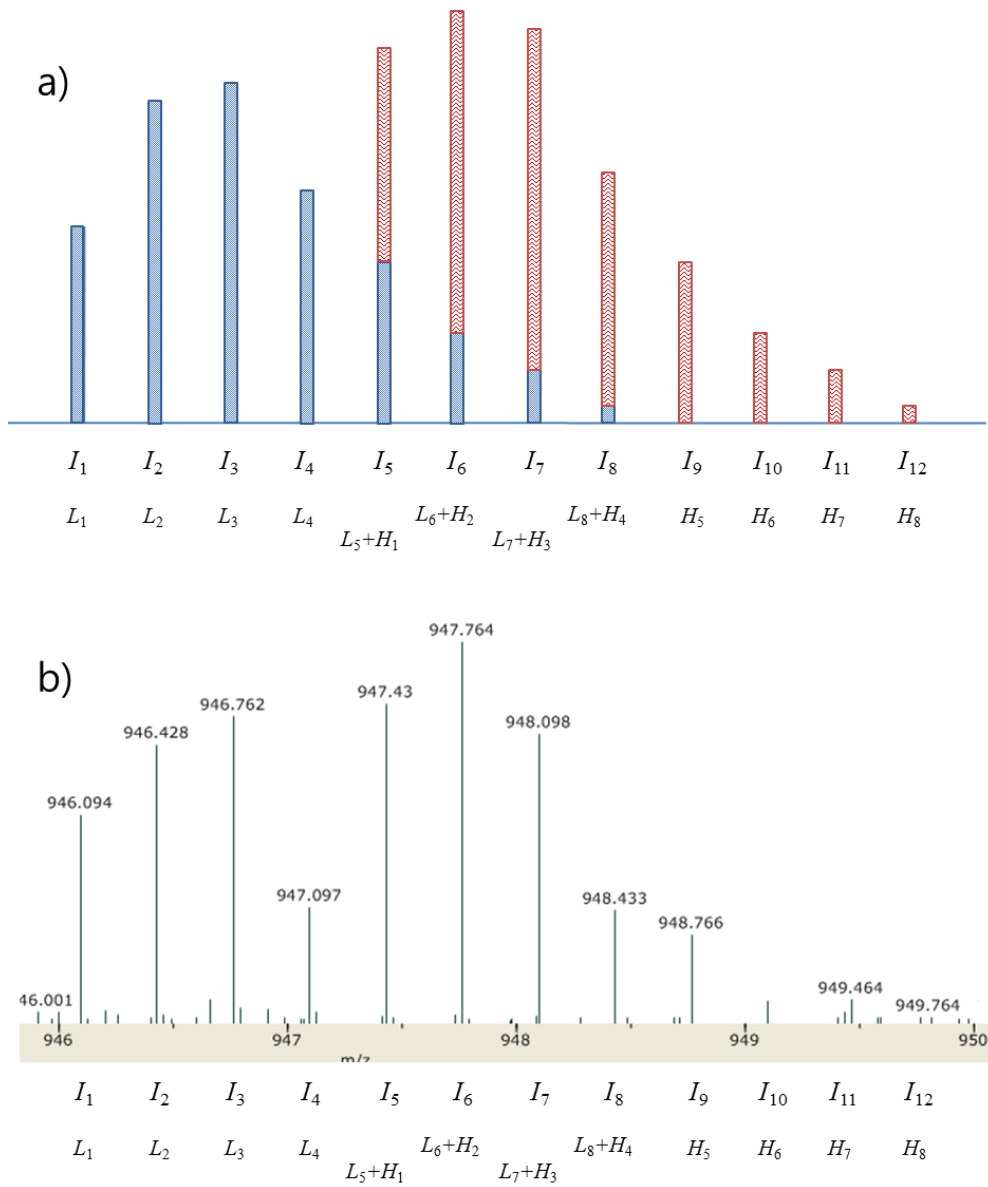


Figure 3.2: Examples of overlapping isotopic clusters. (a) Schematic illustration of overlapping isotopic clusters. (b) Experimental overlapping isotopic clusters of ‘HQPQEFPTYVEPTNDEIcEAFR’ (‘c’ represents a modified cysteine)

Let α be the heavy-to-light (H/L) ratio, i.e. $H_k = \alpha L_k$. Then we can calculate α from I_k values. First, we induce a quadratic equation $I_1\alpha^2 - I_5\alpha + I_9 = 0$ from $I_9 = \alpha L_5$ and $I_5 = L_5 + \alpha L_1$. Using the quadratic formula, we obtain two solutions:

$$\alpha = \frac{I_5 \pm \sqrt{I_5^2 - 4I_1I_9}}{2I_1}.$$

To find an exact solution, we transform them to the equations of L_k :

$$\begin{aligned} \frac{I_5 \pm \sqrt{I_5^2 - 4I_1I_9}}{2I_1} &= \frac{\alpha L_1 + L_5 \pm \sqrt{(\alpha L_1)^2 + 2\alpha L_1 L_5 + L_5^2 - 4\alpha L_1 L_5}}{2L_1} \\ &= \frac{\alpha L_1 + L_5 \pm \sqrt{(\alpha L_1 - L_5)^2}}{2L_1}. \end{aligned}$$

It is easy to see that the larger solution is equal to α if $\alpha \geq L_5/L_1$, and the smaller solution becomes α , otherwise. I_k values are read from the experimental data and L_k can be calculated using the theoretical distribution of the peptide. Therefore, we can calculate α as follows:

$$\alpha = \begin{cases} \frac{I_5 + \sqrt{I_5^2 - 4I_1I_9}}{2I_1} & \text{if } \frac{I_5 + \sqrt{I_5^2 - 4I_1I_9}}{2I_1} \geq \frac{L_5}{L_1}, \\ \frac{I_5 - \sqrt{I_5^2 - 4I_1I_9}}{2I_1} & \text{otherwise.} \end{cases}$$

Similarly, we can induce three more equations from I_k values:

$$I_2\alpha^2 - I_6\alpha + I_{10} = 0,$$

$$I_3\alpha^2 - I_7\alpha + I_{11} = 0,$$

$$I_4\alpha^2 - I_8\alpha + I_{12} = 0.$$

We can calculate multiple α values from these four quadratic equations. Theoretically, all the values should be the same, but the α values calculated using experimental data can be different from each other due to various imperfections in experiments such as low sensitivity, chemical noise, and/or experimental errors. Therefore, it is necessary to integrate the four values. Let α_k be the α value calculated from I_k , I_{k+4} and I_{k+8} , that is,

$$\alpha_k = \frac{H_k + H_{k+4}}{L_k + L_{k+4}}.$$

Since $I_k + I_{k+4} + I_{k+8} = (L_k + L_{k+4}) + (H_k + H_{k+4})$, we get

$$L_k + L_{k+4} = \frac{I_k + I_{k+4} + I_{k+8}}{1 + \alpha_k},$$

$$H_k + H_{k+4} = \frac{\alpha_k(I_k + I_{k+4} + I_{k+8})}{1 + \alpha_k}.$$

By summing these values up, we can calculate the H/L ratio as $\alpha = \sum H_k / \sum L_k$.

Sometimes the quadratic formula for α_k gives no real number solution. This happens when I_k or I_{k+8} are larger than (equivalently, I_{k+4} is smaller than) the theoretically expected intensities. To have at least one real number solution, I_{k+4} must be large enough to satisfy the constraint $I_{k+4}^2 \geq 4I_k I_{k+8}$. Under this assumption, we substitute $2\sqrt{I_k I_{k+8}}$ for I_{k+4} and obtain $\alpha_k = I_{k+4}/2I_k = \sqrt{I_{k+8}/I_k}$ if $I_{k+4}^2 - 4I_k I_{k+8} < 0$.

Extraction of isotopic clusters

For each peptide, we first extract isotopic clusters of the peptide from the precursor MS scan of a tandem MS scan. Because the mass of the peptide and the charge state are obtained from the pepXML file, we can easily locate the first peak of one of the (light or heavy) isotopic clusters. We also find the first peak of the other isotopic cluster depending on the type of label (which is also obtained from the

pepXML file). Subsequently, we extract at most 7 next peaks from each cluster if the two labeled peptides have 8 Da or more mass difference, and we extract at most 11 peaks overall if they have 4 Da mass difference. Each peak is found within 10 ppm mass tolerance.

In experimental data, some peaks from other peptides can be overlapped with the extracted isotopic clusters. To avoid including peaks from these other peptides, which can lead to incorrect quantification, we use the least squares fit values between the extracted isotopic clusters and the theoretical distribution of the peptide. The least squares fit values are calculated as follows:

$$LSF_L = \frac{\sum (T_k - N_L L_k)^2}{\sum T_k^2}, LSF_H = \frac{\sum (T_k - N_H H_k)^2}{\sum T_k^2},$$

where T_k is the relative intensity of the k -th peak in the theoretical distribution of the peptide, and N_L and N_H are normalization factors. The extracted isotopic clusters are used to quantify the ratio of the peptide if at least one of the two least squares fit values is less than 20%.

We also extract isotopic clusters from the scans that are adjacent to the precursor scan of the tandem scan. We provide an automatic determination of elution areas of the precursor ions. First, we consider all scans during 10 seconds from the precursor scan of the tandem scan. Then, we consider subsequent scans until we find two consecutive holes or five holes in total. (A hole is a scan with no isotopic cluster whose least squares fit value is less than 20%.) We also consider scans directly prior to the precursor scan of the tandem scan until we find two consecutive holes or five in total. It is also possible for a user to define a fixed number of scans to be used for quantitation. In this mode, we only consider $\pm n$ MS scans from tandem scan where n is a user-defined number.

Integration of ratios

There may be many scans from which we can calculate a ratio α for a peptide. We need to integrate the ratios obtained from these scans. We consider three integration methods. First, we sum up the intensities for each labeled peptide (for all scans, not only for one scan) and calculate the ratio between summed intensities (called ‘Sum Ratio’). Second, we calculated the weighted average of all the ratios from each scan (called ‘weighted Avg.’). In this case, the sum of intensities in a scan is used as the weight. Third, we calculate a linear regression of the sum of intensities for each labeled peptide using the form of “*Heavy intensity sum* = $\alpha \times$ *light intensity sum*” (called ‘Regression’).

3.2.2 Results

Overview

Human plasma was obtained from healthy volunteers. The six most abundant proteins (serum albumin, immunoglobulin G, immunoglobulin A, transferrin, haptoglobin, and antitrypsin) were depleted using an antibody-based depletion system (MARS column, Agilent Technologies, Palo Alto, CA). The unbound fraction was concentrated using Microcon (3000 Da cutoff, Millipore), and proteins were precipitated by letting stand in 6.5 volumes of cold acetone for 15 min at -20 °C. The precipitate was dissolved in a buffer containing 50 mM Tris-HCl (pH 8.0) and 6 M urea. Protein concentration was determined by the Bradford method.

Two kinds of standard protein mixtures consisted of alpha-lactalbumin, beta-casein, serotransferrin, alpha-S1-casein, alpha-S2-casein, and pancreatic ribonuclease in 50 mM Tris pH 8.0 at different amounts: 10 µg, 10 µg, 20 µg, 25 µg, 25 µg and 10 µg for standard mixture 1 (Std1); 10 µg, 20 µg, 10 µg, 5 µg, 5 µg

sample	light-labeled mixture	heavy-labeled mixture	plasma	mix ratios
S1L1_S2H1	Std1	Std2	No	1:1
S1H1_S2L1	Std2	Std1	No	1:1
S1L1_S2H5	Std1	Std2	No	1:5
S1L5_S2H1	Std1	Std2	No	5:1
PLASMA_S1L1_S2H1	Std1	Std2	Yes	1:1
PLASMA_S1H1_S2L1	Std2	Std1	Yes	1:1

Table 3.1: Sample description for duplex mTRAQ

and 50 μ g for standard mixture 2 (Std2). To experiment on a variety of ratios, we mixed two standard protein mixtures (Std1 and Std2) in various ratios (1:1, 1:5, 5:1). In addition to this, prior to MS analysis, 0.4 mg of the mTRAQ labeled standard protein mixture was added to 1 mg of the trypsin-digested unlabeled plasma proteome in order to test performance under more realistic conditions.

For mTRAQ quantification tests, we prepared seven experimental data sets. We first mixed the same amount of the human plasma samples labeled differently. Because the same sample was used for light and heavy labeling, all the ratios of the peptides identified from this mixture are expected to be 1.0. We also mixed two standard protein mixtures (Std1 and Std2) in various ratios (Table 3.1).

We determined peptide ratios by trying three different methods: ‘Sum Ratio’, ‘Weighted Avg.’ and ‘Regression’. To evaluate the performance of our algorithm, we compared them with XPRESS developed at the Institute for Systems Biology. All programs were executed on the same PC (Intel E6300 processor 1.86GHz, 2GB RAM, Windows XP). For the determination of elution areas of precursor ions, all programs were executed in both automatic mode (AUTO) and user-defined mode with ± 30 scans (FIX). We averaged all $\log_{10}(H/L)$ values of peptides that are

Mode		XPRESS	Sum Ratio	Weighted Avg.	Regression
AUTO	Average $\log_{10}(\text{H/L})$	0.071242	0.034487	0.034569	0.022763
	Standard deviation	0.110822	0.121618	0.151820	0.134611
FIX	Average $\log_{10}(\text{H/L})$	0.070026	0.036780	0.037344	0.031133
	Standard deviation	0.115151	0.102963	0.109182	0.101954

Table 3.2: Performance comparison between different methods
for 1:1 human plasma sample

expected to have the same ratio values, and calculated standard deviations. Because our tests used sampled data with known ratios, the averages and standard deviations of ratios can be a measure to evaluate the correctness of a method. Especially, low standard deviation can be very important for accurate relative quantification. Some peptides have low labeling efficiency because of their chemical properties. In this case, observed ratios can be very different from expected ratios, but they would be considered more reliable if they are all close to a certain value.

1:1 human plasma sample

The same human plasma samples were labeled with light and heavy mTRAQ reagents, respectively. We mixed light and heavy mTRAQ-labeled peptides in the ratio of 1:1. So the average ratios of all peptides were expected to be 1.0. Overall, 2291 peptides were selected by PeptideProphet in this mixture.

The ratios calculated by eight different methods are shown in Table 3.2. The averages of $\log_{10}(\text{H/L})$ values calculated by XPRESS were larger than 0.07 while the largest average value calculated by our method was 0.037344. In this sample, the standard deviations of our method in AUTO mode were larger than XPRESS, but the overall performance of our method was better considering that our average

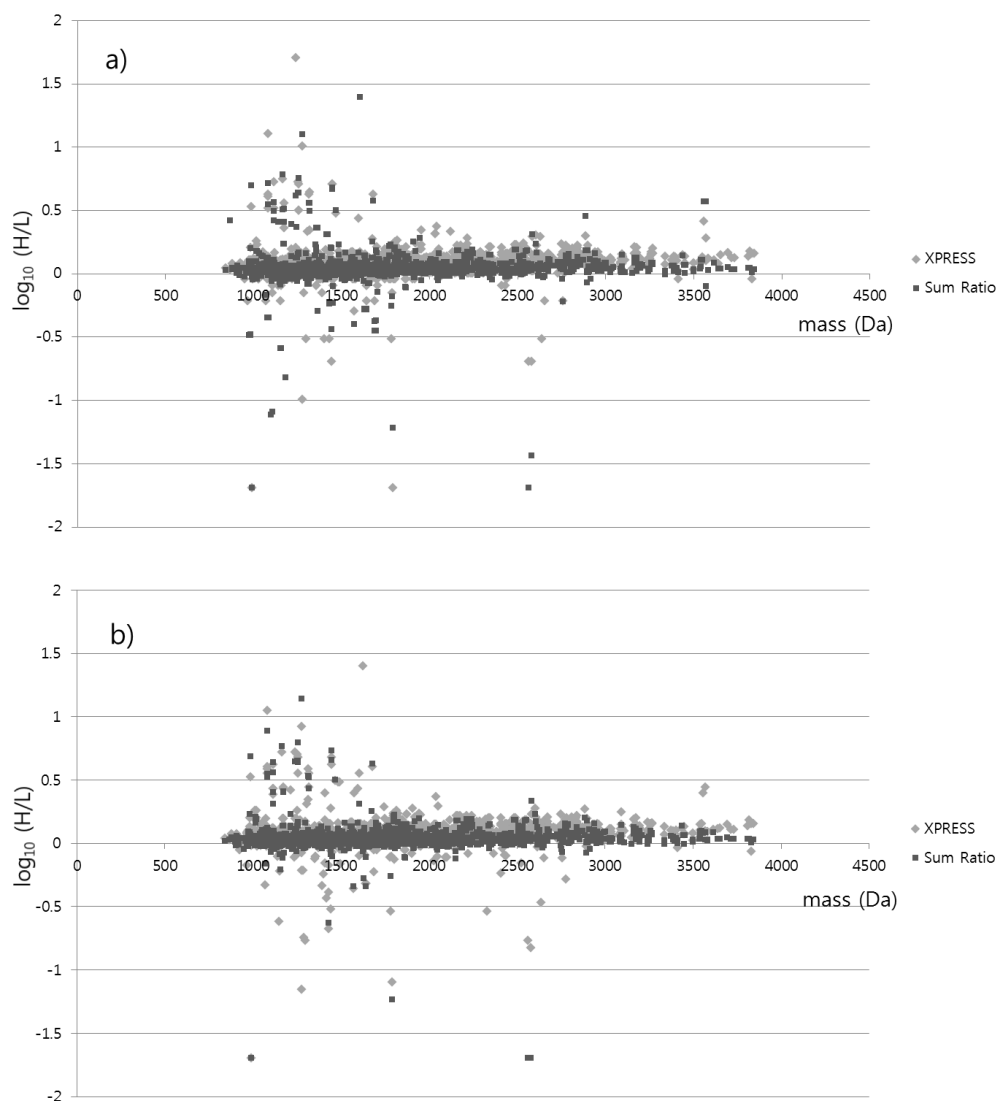


Figure 3.3: Distribution of $\log_{10}(H/L)$ values of peptides from 1:1 human plasma sample. (a) AUTO mode and (b) FIX mode. Though all $\log_{10}(H/L)$ values are expected to be 0.0, most of $\log_{10}(H/L)$ values calculated by XPRESS were biased toward heavy.

was a lot closer to 0.0. All three methods in FIX mode gave more accurate ratios and better standard deviations than XPRESS. In this sample, Regression in FIX mode seemed to give better results, but the differences with other methods were negligible. On the other hand, with the standard mixture experiments, Regression showed worse results than Sum Ratio, especially when expected ratios are far from 1.0. Overall, we recommend using Sum Ratio. Distribution of $\log_{10}(\text{H/L})$ values of XPRESS and Sum Ratio are shown in Figure 3.3. It shows that most of $\log_{10}(\text{H/L})$ values calculated by XPRESS were biased toward heavy.

Ratio of peptides with no lysine

Since mTRAQ is specific to primary amine, the mass difference between heavy- and light-labeled peptides is a multiple of 4 Da depending on the number of Lys residues present, and thus, peptides without any lysine have the smallest mass difference of 4 Da. Our overlap model is especially effective for peptides with no lysine. From 1:1-mixed human plasma sample, we selected the peptides that have no lysine. There were 544 such peptides. Since there are a few outliers whose ratios are far from the expected ratio, it is hard to observe their linearity (Figure 3.4.a-b). Therefore, we removed those outliers whose $\log_{10}(\text{H/L})$ values were larger than 0.5 or smaller than -0.5. Then, we fitted $\log_{10}(\text{H/L})$ values using linear regression (Figure 3.4.c-d). The slope from the results of XPRESS in FIX mode was 6.05862×10^{-5} , which was more than twice larger than that from our algorithm (Sum Ratio in FIX mode), 2.73836×10^{-5} . The ratios calculated by XPRESS were consistently larger than the ratios calculated by our algorithm, especially in high masses. It strongly indicates that our algorithm shows better performance than XPRESS for the quantification of the stable isotope labeled peptides that have an overlap in their isotopic distributions.

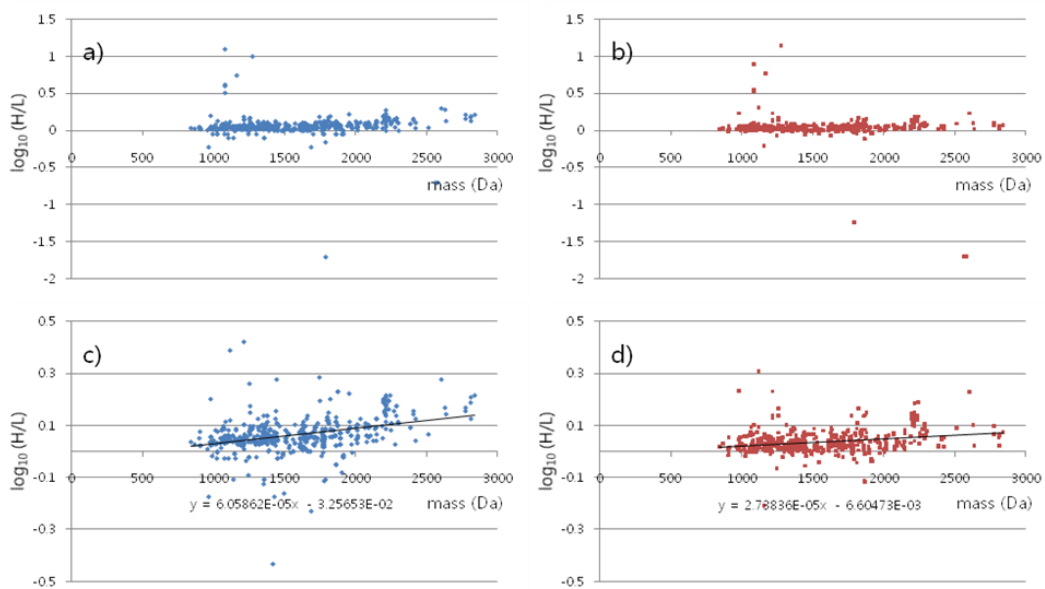


Figure 3.4: Distribution of $\log_{10}(H/L)$ values of peptides with no lysine. (FIX mode)

(a) Distribution of $\log_{10}(H/L)$ values calculated by XPRESS. (b) Distribution of $\log_{10}(H/L)$ values calculated by our algorithm (Sum Ratio). (c) Distribution of $\log_{10}(H/L)$ values calculated by XPRESS after removing outliers (>0.5 or <-0.5). H/L ratios increase as peptide masses increase. (d) Distribution of $\log_{10}(H/L)$ values calculated by our algorithm (Sum Ratio) after removing outliers. The ratios are more consistent than XPRESS regardless of the masses of peptides.

Quantification using isotopic cluster

Our method uses the sum of intensities of detected isotopic cluster for quantification while XPRESS uses only the intensity of the monoisotopic peak. We showed several examples where our method calculated more accurate ratios in Figure 3.5. 'L' and 'H' represent the monoisotopic peak of light- and heavy-labeled peptides, respectively. By checking the existence of isotopic clusters (not just existence of monoisotopic peaks), we can avoid incorrect quantification in various cases.

First, we can exclude peaks of other peptides. In Figure 3.5.a, our method excluded this mass spectrum because Least Square Fit value of light-labeled peptide (LSFL) is 0.31. We can verify that the monoisotopic peak of light-labeled peptide is overlapped with the third peak of a peptide whose mass is 1242.68 Da.

Second, we can quantify accurately even when the monoisotopic peak has abnormal intensity. In Figure 3.5.b, red-dashed lines represent the theoretical isotopic distribution. XPRESS obtained 1.86 as H/L ratio because it uses only the monoisotopic peak. Our method obtained 0.99 as H/L ratio. The relative intensity of monoisotopic peak becomes smaller as the mass of peptide becomes larger, and therefore all peaks of isotopic cluster should be used for better quantification.

Third, we can avoid using the fifth peak of isotopic cluster of light-labeled peptide as heavy-labeled peptide. In Figure 3.5.c, XPRESS used the fifth peak of isotopic cluster of light-labeled peptide to quantify heavy-labeled peptide, and obtained 0.15 as H/L ratio. Our methods only found the isotopic cluster of light-labeled peptide, and obtained 0.02 (user-defined minimum ratio) as H/L ratio. In this case, it seems that peptide was misassigned. The correct mass difference between light and heavy labels should be 8 Da.

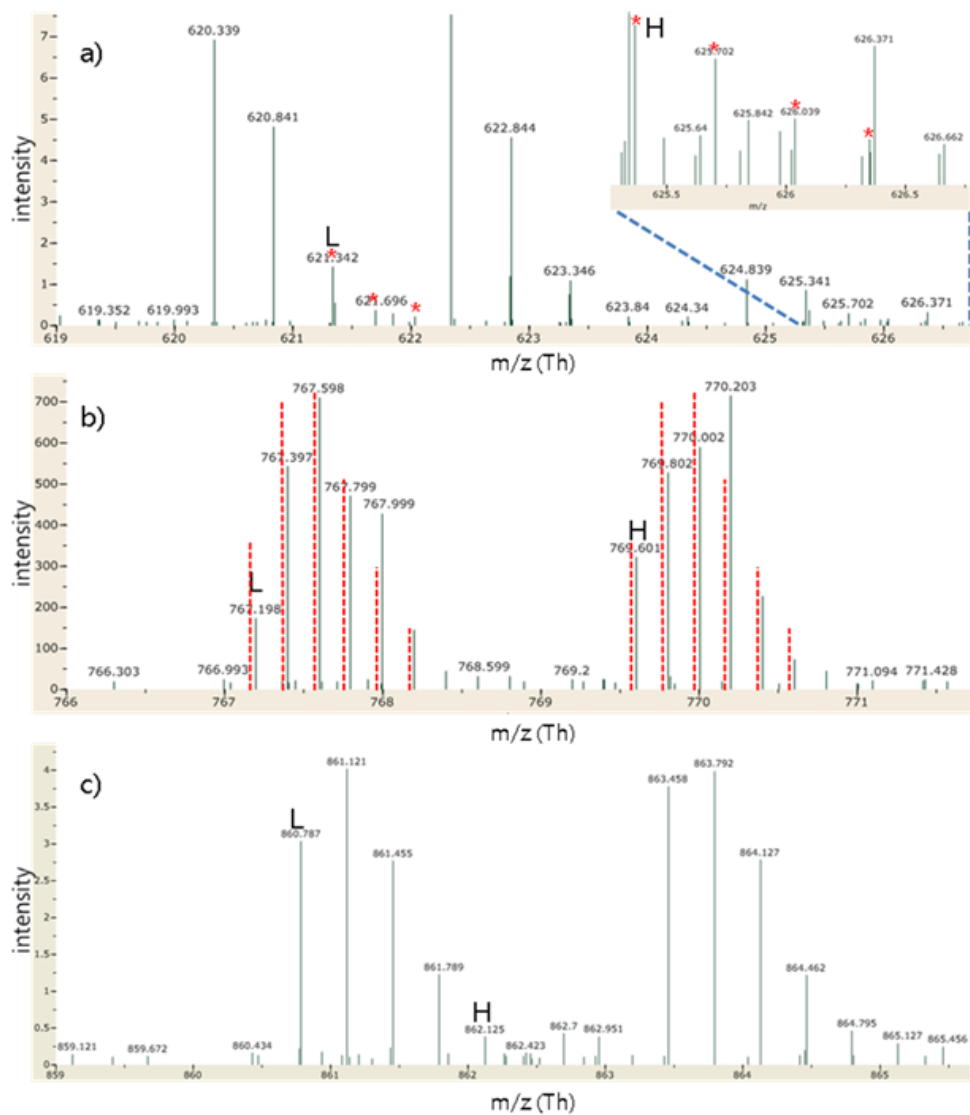


Figure 3.5: Examples where our method calculated more accurate ratios

Standard mixture 1 and Standard mixture 2

For standard protein mixture samples, we present only the results of Sum Ratio method. (Weighted Avg. and Regression methods showed similar but worse ratios and standard deviations.) Expected ratios and computed ratios for each of six proteins are given in Table 3.3-3.8. We first calculated the averages and standard deviations of $\log_{10}(H/L)$ values of peptides. Then, we transformed the averages into H/L scale to compare them to expected ratios. For all mixtures, our method showed similar or better average ratios than XPRESS except pancreatic ribonuclease. Especially, the peptides of Alpha-S1-casein and Alpha-S2-casein showed much better ratios than XPRESS. Furthermore, in spite of inaccurate averages, most of the peptides of pancreatic ribonuclease (except AUTO mode in PLASMA_S1L1_S2H1) showed ratios biased toward Std2. This result implies that there are unknown error factors which make the ratios of peptides of pancreatic ribonuclease incorrectly. Most of standard deviations from our method were also better than XPRESS.

Protein	Expected ratio		XPRESS		Our method (Sum Ratio)	
			AUTO	FIX	AUTO	FIX
Alpha-lactalbumin	1	Average (H/L)	1.230691	1.219947	1.149738	1.137588
		Average Log(H/L)	0.090149	0.086341	0.060599	0.055985
		Standard deviation	0.117583	0.114267	0.079873	0.078114
Beta-casein	2	Average (H/L)	2.006855	2.095975	1.834193	1.920547
		Average Log(H/L)	0.302516	0.321386	0.263445	0.283425
		Standard deviation	0.107177	0.043626	0.113144	0.076320
Serotransferrin	0.5	Average (H/L)	0.538917	0.536943	0.481471	0.470103
		Average Log(H/L)	-0.268478	-0.270072	-0.317430	-0.327807
		Standard deviation	0.088544	0.084211	0.094122	0.076619
Alpha-S1-casein	0.2	Average (H/L)	0.245641	0.246540	0.204650	0.197205
		Average Log(H/L)	-0.609700	-0.608113	-0.688988	-0.705081
		Standard deviation	0.147435	0.123440	0.075113	0.077004
Alpha-S2-casein	0.2	Average (H/L)	0.276269	0.278549	0.237995	0.239169
		Average Log(H/L)	-0.558668	-0.555098	-0.623433	-0.621295
		Standard deviation	0.167923	0.164208	0.159004	0.160128
Pancreatic ribonuclease	5	Average (H/L)	7.580607	7.552296	7.460948	7.698693
		Average Log(H/L)	0.879704	0.878079	0.872794	0.886417
		Standard deviation	0.283208	0.145355	0.126991	0.130309

Table 3.3: Expected ratios and computed ratios in S1L1_S2H1 sample

Protein	Expected ratio		XPRESS		Our method (Sum Ratio)	
			AUTO	FIX	AUTO	FIX
Alpha-lactalbumin	1	Average (H/L)	1.038133	1.044059	1.004826	0.956364
		Average Log(H/L)	0.016253	0.018725	0.002091	-0.019377
		Standard deviation	0.082272	0.088696	0.094697	0.058738
Beta-casein	0.5	Average (H/L)	0.579788	0.586198	0.528163	0.514628
		Average Log(H/L)	-0.236731	-0.231956	-0.277232	-0.288507
		Standard deviation	0.078908	0.065993	0.085222	0.060252
Serotransferrin	2	Average (H/L)	2.380324	2.376561	2.271747	2.269645
		Average Log(H/L)	0.376636	0.375949	0.356360	0.355958
		Standard deviation	0.124123	0.125028	0.099330	0.116415
Alpha-S1-casein	5	Average (H/L)	5.277849	5.069837	4.826430	4.951468
		Average Log(H/L)	0.722457	0.704994	0.683626	0.694734
		Standard deviation	0.312771	0.231996	0.088898	0.101004
Alpha-S2-casein	5	Average (H/L)	4.956373	5.189625	4.759058	4.818879
		Average Log(H/L)	0.695164	0.715136	0.677521	0.682946
		Standard deviation	0.163442	0.088622	0.091010	0.102082
Pancreatic ribonuclease	0.2	Average (H/L)	0.168144	0.165319	0.136372	0.133329
		Average Log(H/L)	-0.774318	-0.781678	-0.865275	-0.875077
		Standard deviation	0.182539	0.175129	0.120305	0.123731

Table 3.4: Expected ratios and computed ratios in S1H1_S2L1 sample

Protein	Expected ratio		XPRESS		Our method (Sum Ratio)	
			AUTO	FIX	AUTO	FIX
Alpha-lactalbumin	5	Average (H/L)	5.755300	5.525073	6.917609	7.053002
		Average Log(H/L)	0.760068	0.742338	0.839956	0.848374
		Standard deviation	0.321660	0.239571	0.078998	0.090011
Beta-casein	10	Average (H/L)	13.714659	12.744682	11.586279	11.883627
		Average Log(H/L)	1.137185	1.105329	1.063944	1.074949
		Standard deviation	0.418943	0.346553	0.129604	0.195382
Serotransferrin	2.5	Average (H/L)	2.994408	2.983878	2.881848	2.869492
		Average Log(H/L)	0.476311	0.474781	0.459671	0.457805
		Standard deviation	0.115859	0.120334	0.068620	0.062601
Alpha-S1-casein	1	Average (H/L)	1.288715	1.288801	1.194904	1.197350
		Average Log(H/L)	0.110157	0.110186	0.077333	0.078221
		Standard deviation	0.054258	0.048935	0.040909	0.050246
Alpha-S2-casein	1	Average (H/L)	1.542403	1.524502	1.387951	1.411569
		Average Log(H/L)	0.188198	0.183128	0.142374	0.149702
		Standard deviation	0.050405	0.060045	0.059325	0.068205
Pancreatic ribonuclease	25	Average (H/L)	51.207079	45.288819	28.540593	34.022091
		Average Log(H/L)	1.709330	1.655991	1.455463	1.531761
		Standard deviation	0.551515	0.506536	0.179232	0.212650

Table 3.5: Expected ratios and computed ratios in S1L1_S2H5 sample

Protein	Expected ratio		XPRESS		Our method (Sum Ratio)	
			AUTO	FIX	AUTO	FIX
Alpha-lactalbumin	0.2	Average (H/L)	0.295901	0.298147	0.255756	0.258905
		Average Log(H/L)	-0.528854	-0.525569	-0.592175	-0.586859
		Standard deviation	0.153596	0.142507	0.125617	0.139198
Beta-casein	0.4	Average (H/L)	0.648700	0.628520	0.566830	0.566405
		Average Log(H/L)	-0.187956	-0.201681	-0.246547	-0.246873
		Standard deviation	0.349004	0.353914	0.362788	0.364027
Serotransferrin	0.1	Average (H/L)	0.147768	0.147968	0.123075	0.119674
		Average Log(H/L)	-0.830421	-0.829833	-0.909831	-0.922000
		Standard deviation	0.340103	0.339036	0.341322	0.337484
Alpha-S1-casein	0.04	Average (H/L)	0.071880	0.075151	0.058639	0.053077
		Average Log(H/L)	-1.143394	-1.124066	-1.231815	-1.275091
		Standard deviation	0.240020	0.213598	0.125220	0.142586
Alpha-S2-casein	0.04	Average (H/L)	0.122571	0.125305	0.063392	0.059175
		Average Log(H/L)	-0.911612	-0.902030	-1.197965	-1.227863
		Standard deviation	0.539069	0.547153	0.243583	0.249255
Pancreatic ribonuclease	1	Average (H/L)	2.546091	2.546719	2.251614	2.307985
		Average Log(H/L)	0.405874	0.405981	0.352494	0.363233
		Standard deviation	0.530181	0.555775	0.404159	0.423759

Table 3.6: Expected ratios and computed ratios in S1L5_S2H1 sample

Protein	Expected ratio		XPRESS		Our method (Sum Ratio)	
			AUTO	FIX	AUTO	FIX
Alpha-lactalbumin	1	Average (H/L)	1.167103	1.173313	1.045890	1.063616
		Average Log(H/L)	0.067109	0.069414	0.019486	0.026785
		Standard deviation	0.071767	0.076176	0.065369	0.069395
Beta-casein	2	Average (H/L)	1.993862	1.988186	1.664604	1.828555
		Average Log(H/L)	0.299695	0.298457	0.221311	0.262108
		Standard deviation	0.040448	0.033685	0.12183	0.051021
Serotransferrin	0.5	Average (H/L)	0.526971	0.529435	0.472549	0.465749
		Average Log(H/L)	-0.278213	-0.276187	-0.325553	-0.331848
		Standard deviation	0.052557	0.065528	0.059061	0.040219
Alpha-S1-casein	0.2	Average (H/L)	0.246812	0.245988	0.194489	0.191135
		Average Log(H/L)	-0.607633	-0.609086	-0.711106	-0.71866
		Standard deviation	0.089503	0.09047	0.038395	0.039765
Alpha-S2-casein	0.2	Average (H/L)	0.278516	0.280265	0.237007	0.239811
		Average Log(H/L)	-0.55515	-0.552431	-0.625239	-0.62013
		Standard deviation	0.058338	0.063346	0.080802	0.081277
Pancreatic ribonuclease	5	Average (H/L)	6.120938	5.973557	3.346162	5.722595
		Average Log(H/L)	0.786818	0.776233	0.524547	0.757593
		Standard deviation	0.159831	0.097629	0.635218	0.099495

Table 3.7: Expected ratios and computed ratios in PLASMA_S1L1_S2H1 sample

Protein	Expected ratio		XPRESS		Our method (Sum Ratio)	
			AUTO	FIX	AUTO	FIX
Alpha-lactalbumin	1	Average (H/L)	1.048682	1.058293	0.958685	0.959951
		Average Log(H/L)	0.020644	0.024606	-0.018324	-0.017751
		Standard deviation	0.066572	0.063705	0.066239	0.066273
Beta-casein	0.5	Average (H/L)	0.613213	0.609960	0.533394	0.545419
		Average Log(H/L)	-0.212389	-0.214699	-0.272952	-0.26327
		Standard deviation	0.032922	0.035318	0.026962	0.028457
Serotransferrin	2	Average (H/L)	2.327148	2.370915	2.230874	2.267410
		Average Log(H/L)	0.366824	0.374916	0.348475	0.35553
		Standard deviation	0.094981	0.102624	0.070709	0.059478
Alpha-S1-casein	5	Average (H/L)	5.595772	5.491525	5.492954	5.480965
		Average Log(H/L)	0.74786	0.739693	0.739806	0.738857
		Standard deviation	0.081621	0.080087	0.071198	0.070827
Alpha-S2-casein	5	Average (H/L)	4.698346	4.739472	4.640534	4.631237
		Average Log(H/L)	0.671945	0.67573	0.666568	0.665697
		Standard deviation	0.080163	0.080312	0.076332	0.078796
Pancreatic ribonuclease	0.2	Average (H/L)	0.225634	0.220806	0.134457	0.178023
		Average Log(H/L)	-0.646596	-0.65599	-0.871417	-0.749525
		Standard deviation	0.16896	0.161123	0.278224	0.134524

Table 3.8: Expected ratios and computed ratios in PLASMA_S1H1_S2L1 sample

3.3 Peptide Quantification Using Triplex mTRAQ

Labeling

3.3.1 Algorithm

Model of overlapping isotopic clusters

We made a schematic model of overlapping triplex isotopic clusters, which is an extension of the model of overlapping duplex isotopic clusters (Figure 3.6). Assuming that an isotopic cluster of a peptide has 8 or less peaks, an overlap exists only if the mass difference between labeled peptides is 4 Da. Therefore, the intensity of the k th peak I_k is given as follows:

$$I_k = \begin{cases} L_k & \text{if } k \leq 4 \\ L_k + M_{k-4} & \text{if } 4 < k \leq n \\ M_{k-4} & \text{if } n < k \leq 8 \\ M_{k-4} + H_{k-8} & \text{if } 8 < k \leq n + 4 \\ H_{k-8} & \text{if } k > n + 4 \end{cases} \quad (1)$$

where n is the number of peaks in the isotopic distribution of a peptide, L_k , M_k , and H_k are the intensities of the k th peaks of the isotopic distributions of the light, medium, and heavy-labeled peptides, respectively.

Let α be the M/L ratio and β be the H/L ratio. For $1 \leq k \leq 4$, it is easy to show

$$L_k = I_{k+4} - I_k \alpha = \frac{I_{k+8} - I_k \beta}{\alpha} = \frac{I_{k+12}}{\beta} \quad (2)$$

from equation (1). Using equation (2), we induced three equations

$$I_k \alpha^2 - I_{k+4} \alpha + I_{k+8} - I_k \beta = 0 \quad (3)$$

$$I_k \beta^2 - I_{k+8} \beta + I_{k+12} \alpha = 0 \quad (4)$$

$$I_k \alpha \beta - I_{k+4} \beta + I_{k+12} = 0 \quad (5)$$

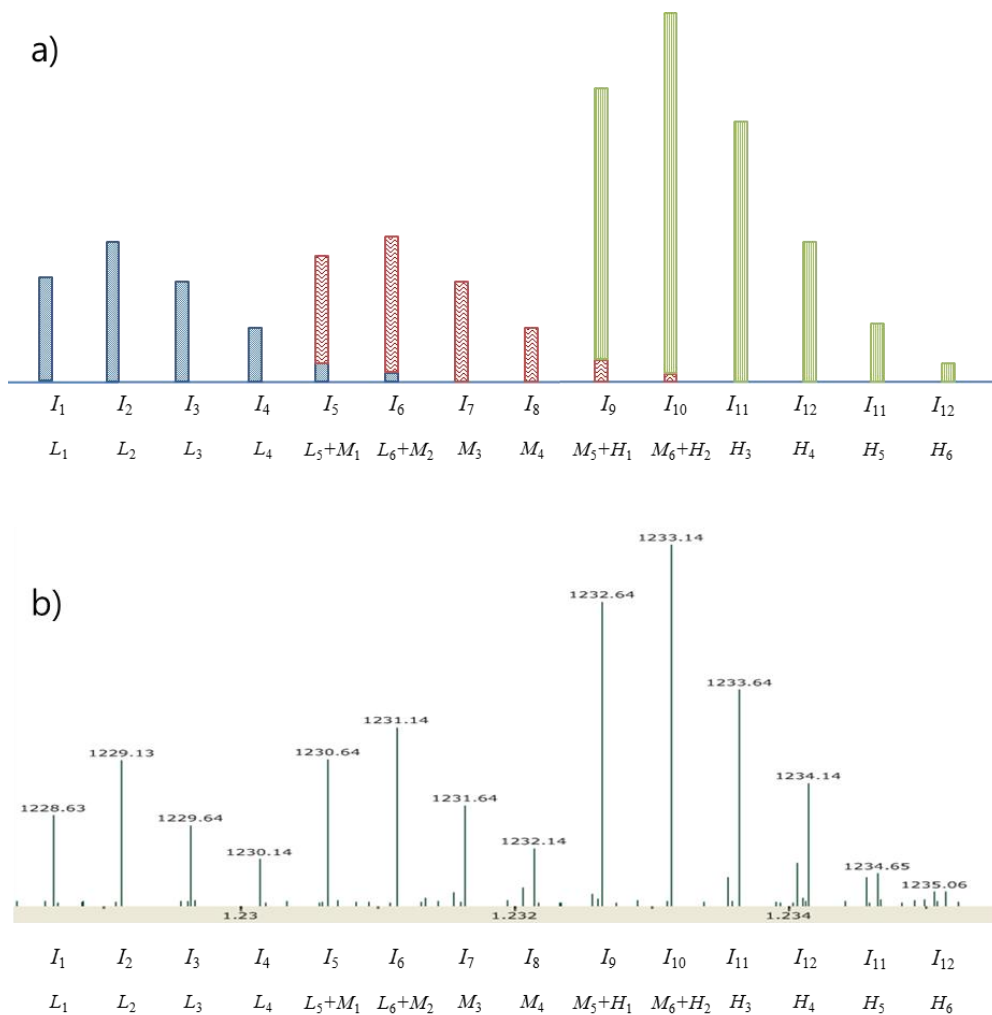


Figure 3.6: Examples of overlapping triplex isotopic clusters

(a) Schematic model of overlapping triplex isotopic clusters

(b) Experimental overlapping isotopic clusters of 'EPMIGV NQELAYFYPELFR'

From equations (4) and (5), we obtain a cubic equation for β :

$$I_k^2 \beta^3 - I_k I_{k+8} \beta^2 + I_{k+4} I_{k+12} \beta - I_{k+12}^2 = 0 \quad (6)$$

Solving equation (6), we obtain up to three candidate values for β . Then, by substituting the candidates into equation (3) and solving it, we obtain up to two candidate values for α . (Substituting candidates for β into equation (4) may lead to an abnormal α value because I_{k+12} could possibly be very small and inaccurate in its value. Substituting into Equation (5) could also be problematic because a low β value could cause an inaccurate α value.) To select the most accurate ratio pair, we define an error value:

$$\text{Err} = \left(\frac{T_{k+4}}{T_k} - \frac{L_{k+4}}{L_k} \right)^2 + \left(\frac{T_{k+4}}{T_k} - \frac{M_{k+4}}{M_k} \right)^2 + \left(\frac{T_{k+4}}{T_k} - \frac{H_{k+4}}{H_k} \right)^2 \quad (7)$$

where T_k is the intensity of the k th peak of the theoretical isotopic distribution of the peptide. (The EMASS algorithm was used to calculate T_k values [31].) The error value should be very small for the correct ratio pair because L_{k+4}/L_k , M_{k+4}/M_k , and H_{k+4}/H_k are theoretically the same as T_{k+4}/T_k . Therefore, we calculated the error value for each candidate pair and select the pair with the lowest error value. After all pairs for $1 \leq k \leq 4$ are selected, we can calculate the M/L ratio $\alpha = \sum M_k/L_k$ and the H/L ratio $\beta = \sum H_k/L_k$.

Determination of the elution areas of peptides

In most LC/MS experiments, tandem MS scans are acquired using dynamic exclusion (DE). For each MS/MS scan, therefore, we know only one MS scan where the identified peptide is eluted. We need to determine the elution area of the peptide because it is eluted over a period of time. However, some peptides have similar atomic masses and elution times, so their elution areas can have overlaps. A naive approach such as using a fixed range (e.g. within ± 30 s from the tandem scan

of peptides) has a risk to use incorrect MS scans where other peptides are overlapped. Therefore, it is very important to determine accurate elution areas of the peptides for accurate relative quantification.

We assume that the distribution of peptide elution time is approximately a normal distribution. Because of noise and overlap of peptides, MS scans with low intensities at both ends of the elution area may not be trusted. If we use only MS scans with high total ion current while modeling the elution profile as a normal distribution, the mean μ of the normal distribution can be approximated, but the variance σ^2 can't. Instead, we use the full width at half maximum (FWHM) to induce σ^2 . From the probability density function of the normal distribution, we deduce $e^{-\frac{(FWHM/2)^2}{2\sigma^2}} = 1/2$ and obtain $\sigma^2 = FWHM^2/8 \ln 2$.

When a peptide identification and the associated tandem MS scan is given, our algorithm first finds the maximum point of the peptide's elution profile. For each MS scan within $\pm 30s$ range from the given tandem scan, it identifies triplex isotopic clusters and calculates the sum of intensities. (Details are explained in the next section.) The MS scan whose sum of intensities is the highest is selected as the maximum point of the elution area. Then it extends the elution area while the sum of intensities of MS scan is above a half of that of the maximum point. The length of the extended area is used as FWHM and weighted average time of scans in the extended area is used as μ . The area with higher intensities than 10% of the maximum intensity in the normal distribution (from $\mu - \sqrt{FWHM^2 \ln 10 / 4 \ln 2}$ to $\mu + \sqrt{FWHM^2 \ln 10 / 4 \ln 2}$) is used as the elution area of a peptide. As an example, the elution area approximation for 'HPIKHQGLPQEVLENLLR' is shown in Figure 3.7. From the given elution time (951.423 s), where tandem MS scan was acquired, we first found the maximum point of elution area (957.93 s).

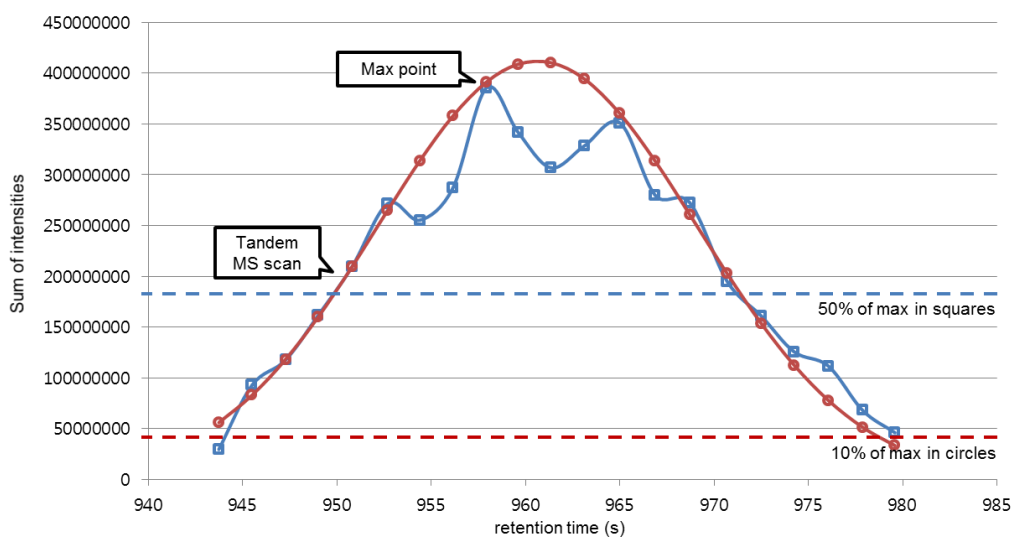


Figure 3.7: Elution area approximation to normal distribution

Elution area approximation for 'HPIKHQGLPQEVLNENLLR'. The line with squares represents the sum of intensities of the peptide over the elution area and the line with circles is an approximated normal distribution.

Then we extended the area until the sum of intensities is below 50% of that of the maximum point and obtained $\mu = 960.63$ and $\text{FWHM} = 19.9$. Finally, we used the area with higher intensities than 10% of the maximum intensity of the approximated normal distribution.

Our algorithm calculates M/L and H/L ratios for all MS scans in the elution area. Then, each of the set of M/L and H/L ratios is integrated by linear regression using the form “ $y = cx$ ”. The intensities of peaks are split into the intensity of light-, medium-, and heavy-labeled peptide. We estimate c using the set of intensities of light-labeled peptides as x_i ’s, and the set of intensities of medium- and heavy-labeled peptides as y_i ’s for M/L and H/L ratios, respectively.

Identification and validation of triplex isotopic clusters

For each MS scan in the elution area, our algorithm identifies isotopic clusters of a target peptide. Let MZ_k be the m/z of the k th peak of an isotopic cluster, then we can calculate three MZ_1 ’s corresponding to triplex isotopic clusters from the given sequence, charge z , and modification. Our algorithm first finds the monoisotopic peak of each isotopic cluster from MZ_1 within 10 ppm error tolerance. Then, it finds subsequent isotopic peaks from $MZ_k = MZ_{k-1} + 1.00235/z$ within 10 ppm error tolerance. The k th peak is inserted to the isotopic cluster only if the peak improves the least squares fit value (LSQ). If the LSQ between the theoretical distribution of the peptide and the isotopic cluster without the k th peak is lower than that with the k th peak, the k th peak is discarded and it doesn’t look for any more peaks. If there are two or more candidate peaks for the k th peak, the peak with the lowest LSQ is selected. For example, there are two candidates for the second isotopic peak of the heavy-labeled isotopic cluster and the smaller peak is selected in Figure 3.8.a.

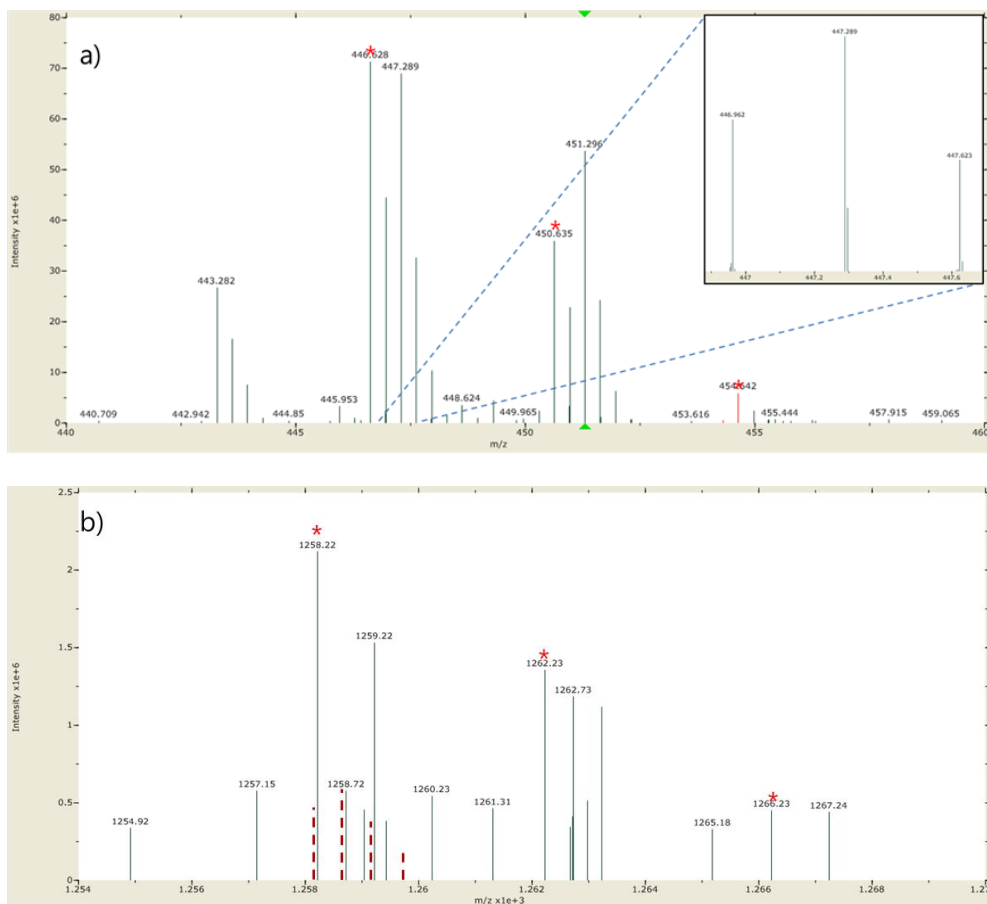
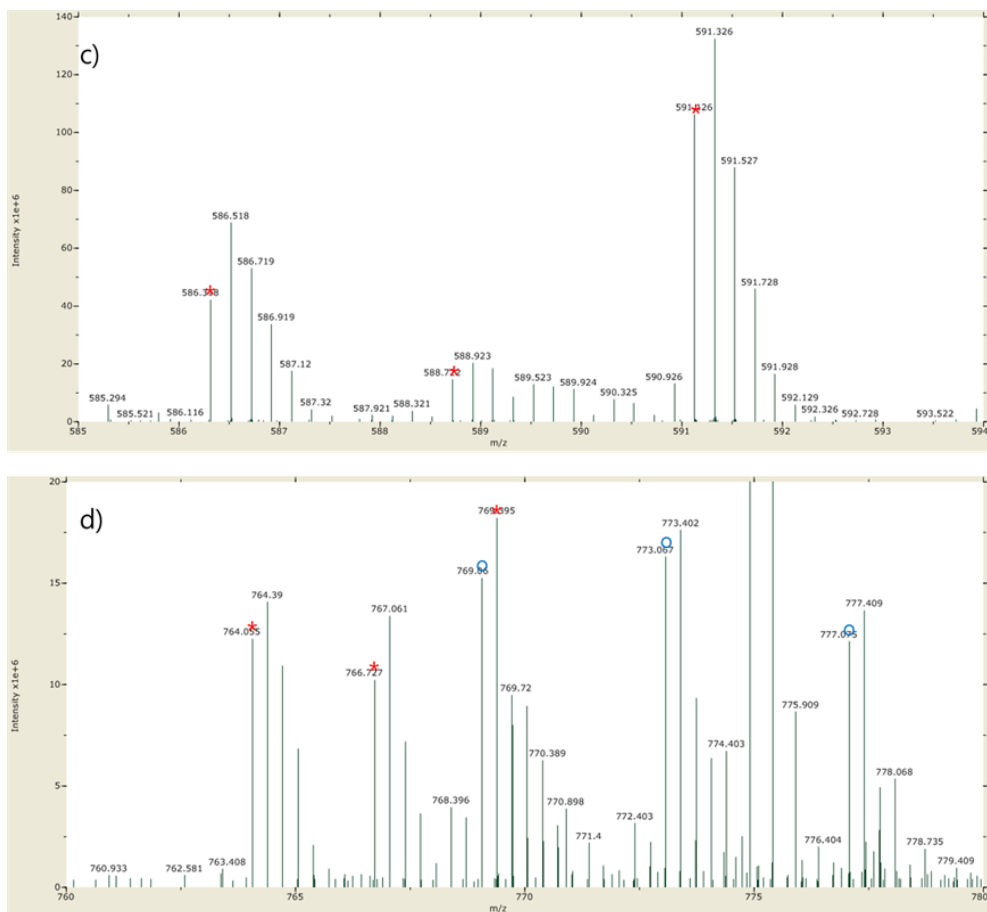


Figure 3.8: Four types of overlaps between chemically different peptides

(a) Two isotopic clusters are overlapped, but no isotopic peak is shared.

(b) An MS scan in the elution area of ‘HPIKHQGLPQEVLNENLLR’ of which expected ratio is 3:1:1. The dashed lines represent its theoretical isotopic distribution. Since an isotopic cluster with a different charge value is overlapped with the light-labeled isotopic cluster, the LSQ value becomes significantly high, so we discard this MS scan during the quantification of the target peptide.



(c) An MS scan in the elution area of ‘TVGGKEDVIWELLNHAQEHLFGK’ of which expected ratio is 3:1:10. An isotopic cluster with the same charge and a higher mass is overlapped with the medium-labeled isotopic cluster. Since the fifth peak increases the LSQ value, only the first to the fourth peaks are used to quantify.

(d) An MS scan in the elution area of ‘GITWGEETLMEYLENPK’ of which expected ratio is 3:3:1. An isotopic cluster with the same charge and 1 Da smaller mass is overlapped with the heavy-labeled isotopic cluster. Since it is difficult to separate these overlapping isotopic clusters accurately, we discard this MS scan during the quantification.

After identification of triplex isotopic clusters of a target peptide, we check them and discard the current MS scan if they are doubtful according to the following criteria. First, we check whether the overall shape of each isotopic cluster resembles that of a theoretical isotopic distribution. At least the LSQ of the most abundant isotopic cluster must be below a threshold (e.g. 0.2). The LSQ of the others also should be below the threshold unless their sums of intensities are lower than a half of that of the most abundant isotopic cluster. (If an isotopic cluster has low abundance, its shape could be abnormal because it may be interfered by chemical noise and other peptides.) Second, we check whether the identified isotopic cluster is overlapped with another peptide. Four types of overlaps are shown in Figure 3.8. There is no problem if no isotopic peak is shared by two isotopic clusters (Figure 3.8.a). If an isotopic cluster with a different charge value is overlapped, the LSQ of the identified isotopic cluster should be significantly high, so we can discard the current MS scan (Figure 3.8.b). If an isotopic cluster with the same charge and a higher mass is overlapped, shared isotopic peaks could not be inserted to the isotopic cluster of the target peptide because it increases the LSQ of the isotopic cluster (Figure 3.8.c). Only the case in which an isotopic cluster with the same charge and a lower mass is overlapped needs additional filtering (Figure 3.8.d). We can easily detect these overlaps by considering previous peaks, but we can't separate overlapping isotopic clusters in this case because they look like one isotopic cluster. Therefore, we discard the current MS scan if at least one isotopic cluster of a target peptide could be identified as an isotopic cluster with the same charge and a lower mass.

Protein	Std1 (μg)	Std2 (μg)	Std3 (μg)
alpha-lactalbumin (LALBA)	5	5	5
beta-casein (CSN2)	5	10	1
Serotransferrin (TF)	10	1	3
alpha-S1-casein (CSN1S1)	1	1	3
alpha-S2-casein (CSN1S2)	1	1	3
cytochrome c (CYCS)	3	3	1
beta-lactoglobulin (LGB)	1	5	10
Total	26	26	26

Table 3.9: Standard protein mixtures for triplex mTRAQ

3.2.2 Results

Application to 7-standard protein data mixed with known ratios

We analyzed two datasets in which seven standard proteins were mixed in different ratios. For the Set1 experiment, Std1 was labeled with light, Std2 with medium, and Std3 with heavy. For the Set2 experiment, Std1 was labeled with heavy, Std2 with medium, and Std3 with light. The expected ratios for each experiment are shown in Table 3.9.

After validation, we obtained 147 MS/MS scans from Set1 and 139 MS/MS scans from Set2 (168 unique peptides). We calculated M/L and H/L ratios of the peptides and classified them according to the proteins. Then we calculated the averages of ratios in individual cases and compared them to the expected ratios (Table 3.9). The M/L ratios were generally similar to the expected ratios except CSN2 and CSN1S2, whose ratios were somewhat higher than expected ratios. Most H/L ratios were somewhat lower than expected ratios, but their standard

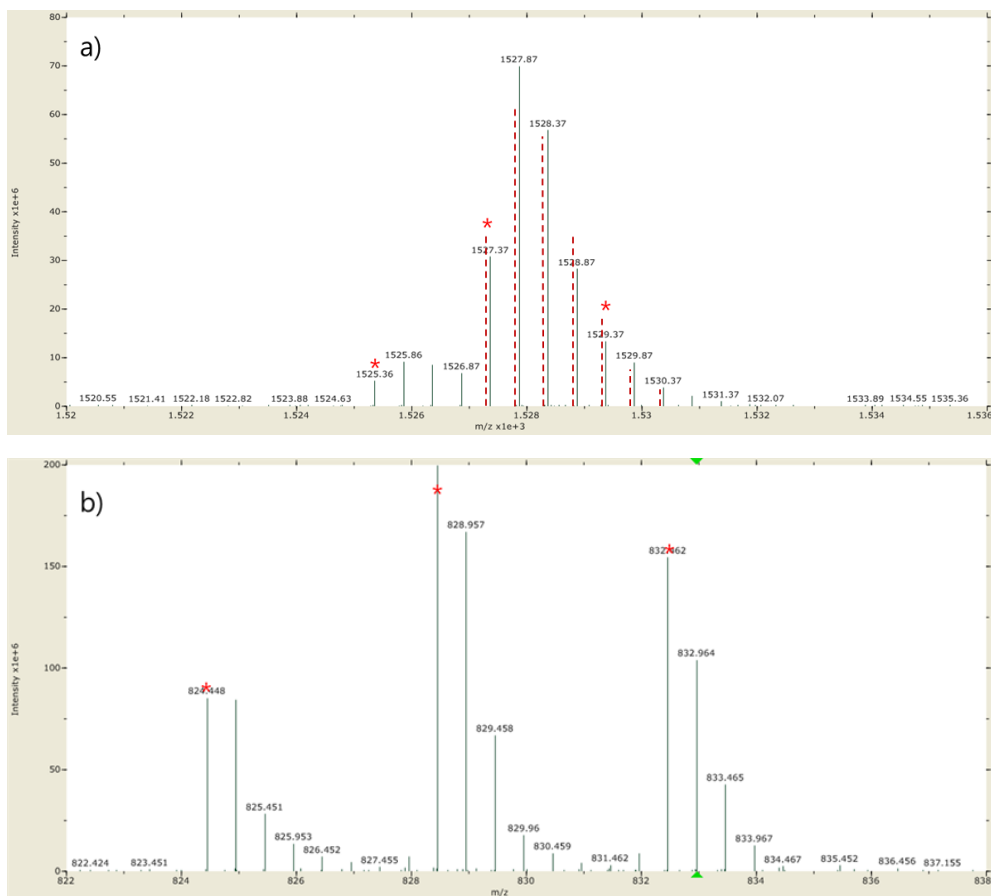


Figure 3.9: Manual inspection for the peptides whose computed ratios are different from the expected ratio

(a) The most abundant isotopic clusters for 'DMPIQAFLLYQEPVLGPVRGPFPIIV'. The dashed lines represent its theoretical isotopic distribution. The expected ratio is 5:10:1, and our algorithm computed 5.612575 as M/L and 0.163601 as H/L for this peptide.

(b) The most abundant isotopic clusters for 'ALNEINQFYQK'. The expected ratio is 1:1:3, and our algorithm computed 2.074808 as M/L and 1.48508 as H/L. It is clear that our ratios are more suitable than the expected ratios for these isotopic clusters.

deviations are meaningfully small. We manually inspected the isotopic clusters of these peptides and concluded that the computed ratios are certainly correct despite their discrepancy from the expected ratios. Some examples of these cases are shown in Figure 3.9. In spite of our effort to label the samples and to mix them accurately, the mixed ratios of samples may be very different from the expected ratio because of low labeling efficiency, chemical property, and experimental error. However, the ratios of peptides of the same protein should be always similar, so low standard deviations give strong evidence that our computed ratios were accurately determined. Figure 3.10 shows the distribution of ratios for LALBA. Each of M/L and H/L ratios represent similar values.

Cause of low abundance of heavy-labeled peptides

Std1 and Std3 were labeled Light and Heavy mTRAQ, respectively, in Set1 Experiment and vice-versa in Set2 Experiment. The calculated H/L ratios were lower than the estimated values in both cases, which exclude the possibility of under-digestion of some of the standard mixtures. If so, we would expect reversed H/L ratios between the two experimental sets. It becomes even more evident if we consider the MS2 search results in which only one out of 168 validated peptides was identified as partially labeled.

The root cause can also be explained, though in part, by isotope impurity of heavy label. Upon closer inspection of MS1 spectra of identified peptides, a peak 1 Da smaller than the monoisotopic peak of heavy label was frequently found. It was reported that iTRAQ reagents contain trace levels of isotopic impurities. Since mTRAQ shares the same chemical structure with iTRAQ, we expect that the same problem will happen in mTRAQ data analysis.

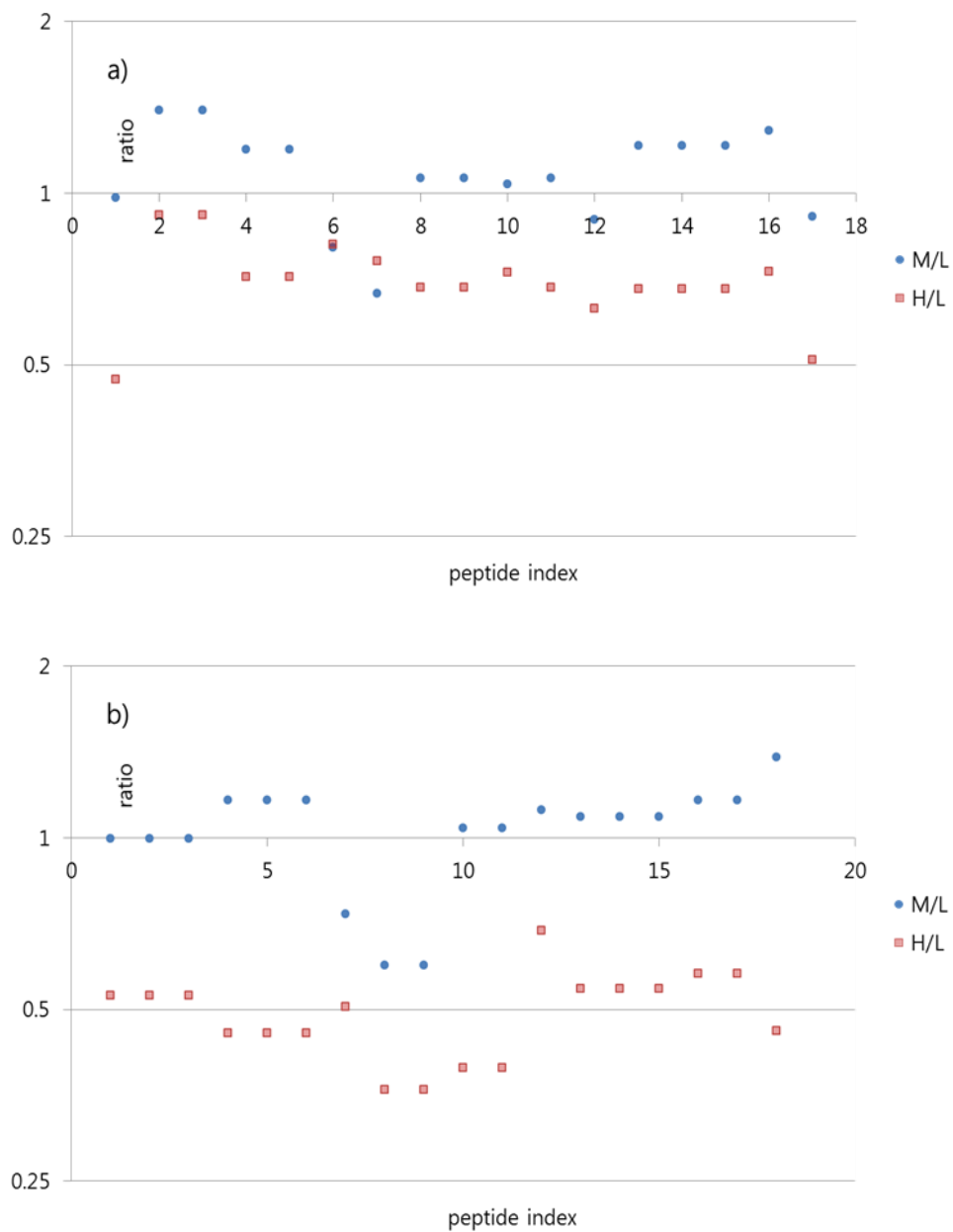


Figure 3.10: Distribution of ratios of peptides for LALBA

(a) Set1 experiment, (b) Set2 experiment

Another possibility is low labeling efficiency of the heavy reagent. If we assume that the M/L ratios are correct, we can approximate the H/L ratios in Set1 experiment using M/L ratios in Set2 experiment. Similarly, we can approximate the H/L ratios in Set2 experiment using M/L ratios in Set1 experiment. We compared them with the computed H/L ratios and observed that the computed H/L ratios are consistently 50~70% of the approximated H/L ratios except for the cases of CYCS in Set1 experiment (Table 3.10). This result shows the possibility that the heavy reagent had low labeling efficiency.

Protein	Number of MS/MS	M/L			H/L		
		Expected ratio	Our ratio	Standard deviation	Expected ratio	Our ratio	Standard deviation
LALBA	17	1	1.074858	0.085456	1	0.695681	0.072863
CSN2	3	2	4.847636	0.093352	0.2	0.184493	0.045277
TF	76	0.1	0.098397	0.186558	0.3	0.160631	0.141608
CSN1S1	10	1	1.264178	0.10274	3	1.948002	0.085701
CSN1S2	4	1	2.419368	0.133448	3	1.644656	0.088651
CYCS	15	1	0.846116	0.068538	0.3	0.34976	0.123581
LGB	22	5	5.138141	0.181286	10	8.143161	0.174779

(a) Set1 experiment

Protein	Number of MS/MS	M/L			H/L		
		Expected ratio	Our ratio	Standard deviation	Expected ratio	Our ratio	Standard deviation
LALBA	18	1	1.010655	0.098869	1	0.487596	0.076048
CSN2	3	10	13.64935	0.112667	5	1.517403	0.112521
TF	64	0.33	0.43843	0.166834	3.3	2.829079	0.108205
CSN1S1	13	0.33	0.471544	0.071577	0.33	0.223681	0.084549
CSN1S2	10	0.33	0.984124	0.042867	0.33	0.274309	0.083823
CYCS	13	3	1.83768	0.049094	3	1.417869	0.083364
LGB	18	0.5	0.436385	0.049273	0.1	0.050817	0.220646

(b) Set2 experiment

Table 3.10: Expected ratios and computed ratios for seven proteins in standard mixtures

Chapter 4

Conclusion

In this thesis, we have considered two problems which we encountered while analyzing mass spectrometry data, and have presented the following three results to improve both throughput and quality of analysis.

First, we have presented a new probabilistic model for isotopic distributions and a novel algorithm RAPID for determining isotopic distributions and monoisotopic masses based on the model. RAPID was applied to protein mixture data from a high resolution mass spectrometer and we obtained better performance than those of THRASH-based implementations [7]. RAPID found more isotopic clusters of identified peptides in spite of the similar number of the total clusters. It does not use the *averagine* [5] fitting method, so we successfully resolve the 1-2Da mismatch problem in THRASH, which occurs especially on isotopic clusters that deviate from the *averagine* distribution due to their weak intensity. Overlapping clusters are also identified successfully in RAPID. Because RAPID uses simple ratio functions to evaluate the score of isotopic clusters, its execution time is very fast. This speed is expected to allow “on-the-fly” determination of monoisotopic

masses during an LC/MS/MS experiment, which provides advantages such as accurate assignment of precursor monoisotopic masses to the corresponding MS/MS data. Shah et al. used RAPID in their on-line data analysis system for fast execution [45]. Several other researchers also introduced it into their experimental frameworks [46-49], and the new model is being put to use in the modeling and analysis of several specific proteins [50, 51]. RAPID is integrated into Decon2LS as an alternative of THRASH, and it is available on <http://omics.pnl.gov/software/decontools-decon2ls> [52].

Second, we have presented a new algorithm QuadQuant for the peptide quantification in duplex mTRAQ labeling experiments that can overcome errors resulting from overlapping isotopic clusters of heavy- and light-labeled peptides. Using the quadratic equations induced from the theoretical distribution model of overlapping isotopic clusters, our algorithm could separate the overlapping isotopic clusters and quantify the ratio of isotope labeled peptides more accurately and reliably than an existing method XPRESS [18], especially for the peptides whose mass difference between labels is relatively small. We expect that this algorithm can be extended to other labeling methods such as ^{18}O labeling. Because the mass shift in ^{18}O labeling is also 2 or 4 Da, it can be analyzed based on the same principle we used for mTRAQ labeled peptides without lysine. QuadQuant obtained good averages and standard deviations for the peptides whose expected ratios lie between 0.1 and 10.0, but failed to obtain good results for peptides with bigger differences in quantity, similarly with an existing method. It seems that the current methodology is not sensitive enough to handle proteins whose quantitative difference exceeds more than an order of magnitude. It has also been contributing in further researches [46, 53]. We exploited high mass accuracy and high resolution of Orbitrap mass spectrometer, and successfully corrected the biases introduced by

overlapped isotope clusters. But further study is required both in terms of mass spectrometry and data analysis.

- In mass spectrometry experiment, we need to explore various other ways we can further improve sensitivity in quantitation. For instance, we can improve chromatographic conditions by introducing multi-dimensional separation; we can try out different modes of operation for data acquisition by mass spectrometers such as making use of inclusion/exclusion lists.
- Because our algorithm used the peptide sequence assigned by the database search software such as SEQUEST, it may extract a wrong isotopic cluster as the pair of the isotopic cluster that corresponds to the identified peptide if the search engine assigned an incorrect peptide sequence to the tandem MS of the peptide. To overcome such a problem, different search strategies may be needed, for example, checking all other isotopic clusters that could be paired with the isotopic cluster obtained from peptide mass. However, it leads to a difficult problem (e.g. the peptide ID correction problem) that warrants further research.
- Our algorithm doesn't consider the theoretical isotopic distribution, so it could lead to inaccurate results for the peptides which are overlapped with other peptides. It might be improved by adjusting ratios using the theoretical overlapping isotopic distributions, but we leave it as a future work. By the way, most of other analysis tools are focusing on their overall framework, the techniques for specific labeling method, the estimation of protein ratios using peptide ratios, and the determination of elution areas. There are a few programs which consider the overlap of isotopic clusters of labeled peptides [24, 28]. And it is hard to compare

our algorithm with these programs due to the difference of target labeling methods. IEMM [28] introduced a Gaussian model for overlapping isotopic clusters on ^{18}O labeling. It would be a meaningful challenge to apply this model to mTRAQ labeling and compare with our algorithm.

Finally, we have presented a new algorithm for peptide quantification in triplex mTRAQ experiments. It is an extension of our algorithm for duplex mTRAQ experiments and can calculate the ratios of peptides accurately by separating overlapping triplex isotopic clusters based on the arithmetic models of isotope overlap. It also includes an automatic determination for the elution area of peptides. Use of intensity information considering the shape of an elution curve of a peptide improved the accuracy of our method. When used within the TPP pipeline, it can easily analyze high-throughput proteomics data.

Bibliography

- [1] H. D. Niall. Automated Edman degradation: the protein sequenator. *Methods Enzymol* **27**: 942-1010, 1973.
- [2] J. B. Fenn, M. Mann, C. K. Meng, and S. F. Wong. Electrospray Ionization-Principles and Practice. *Mass Spectrometry Reviews* , **9**(1): 37-70, 1990.
- [3] M. Karas and F. Hillenkamp. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10000 Daltons. *Analytical Chemistry*, **60**(20): 2299-2301, 1988.
- [4] M. Mann, C. K. Meng, and J. B. Fenn. Interpreting Mass-Spectra of Multiply Charged Ions. *Analytical Chemistry*, **61**(15): 1702-1708, 1989.
- [5] M. W. Senko, S. C. Beu, and F. W. McLafferty. Automated Assignment of Charge States from Resolved Isotopic Peaks for Multiply-Charged Ions. *Journal of the American Society for Mass Spectrometry*, **6**(1): 52-56, 1995.
- [6] Z. Zhang and A. G. Marshall. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Journal of the American Society for Mass Spectrometry*, **9**(3): 225-233, 1998.

- [7] D. M. Horn, R. A. Zubarev, and F. W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11(4), 320-332, 2000.
- [8] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422, (6928), 198-207, 2003.
- [9] M. Mann and M. Wilm. Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry*, 66(24), 4390-4399, 1994.
- [10] J. K. Eng, A. L. McCormack, and J. R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11), 976-989, 1994.
- [11] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18), 3551-3567, 1999.
- [12] L. N. Mueller, M. Y. Brusniak, D. R. Mani, and R. Aebersold. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *Journal of Proteome Research*, 7(1), 51-61, 2008.
- [13] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17(10), 994-999, 1999.
- [14] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell

- culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1(5), 376-386, 2002.
- [15] X. D. Yao, A. Freas, J. Ramirez, P. A. Demirev, and C. Fenselau. Proteolytic O-18 labeling for comparative proteomics: Model studies with two serotypes of adenovirus. *Analytical Chemistry*, 73(13), 2836-2842, 2001.
- [16] I. I. Stewart, T. Thomson, and D. Figeys. O-18 Labeling: a tool for proteomics. *Rapid Communications in Mass Spectrometry*, 15(24), 2456-2465, 2001.
- [17] L. V. DeSouza, A. M. Taylor, W. Li, M. S. Minkoff, A. D. Romaschin, T. J. Colgan, and K. W. M. Siu. Multiple reaction monitoring of mTRAQ-labeled peptides enables absolute quantification of endogenous levels of a potential cancer marker in cancerous and normal endometrial tissues. *Journal of Proteome Research*, 7(8), 3525-3534, 2008.
- [18] D. K. Han, J. Eng, H. L. Zhou, and R. Aebersold. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology*, 19(10), 946-951, 2001.
- [19] X. J. Li, H. Zhang, J. A. Ranish, and R. Aebersold. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Analytical Chemistry*, 75(23), 6648-6657, 2003.
- [20] T. Shinkawa, M. Taoka, Y. Yamauchi, T. Ichimura, H. Kaji, N. Takahashi, and T. Isobe. STEM: A software tool for large-scale proteomic data analyses. *Journal of Proteome Research*, 4(5), 1826-1831, 2005.

- [21] B. D. Halligan, R. Y. Slyper, S. N. Twigger, W. Hicks, M. Olivie, and A. S. Greene. ZoomQuant: An application for the quantitation of stable isotope labeled peptides. *Journal of the American Society for Mass Spectrometry*, 16(3), 302-306, 2005.
- [22] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. W. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. McIntosh. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, 22(15):1902-1909, 2006.
- [23] W. T. Lin, W. N. Hung, Y. H. Yian, K. P. Wu, C. L. Han, Y. R. Chen, Y. J. Chen, T. Y. Sung, and W. L. Hsu. Multi-Q: a fully automated tool for multiplexed protein quantitation. *Journal of Proteome Research*, 5(9):2328-2338, 2006.
- [24] V. Faca, M. Coram, D. Phanstiel, V. Glukhova, Q. Zhang, M. Fitzgibbon, M. McIntosh, and S. Hanash. Quantitative analysis of acrylamide labeled serum proteins by LC-MS/MS. *Journal of Proteome Research*, 5(8):2009-2018, 2006.
- [25] M. E. Monroe, N. Tolic, N. Jaitly, J. L. Shaw, J. N. Adkins, and R. D. Smith, VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics*, 23(15):2021-2023, 2007.
- [26] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), 1367-1372, 2008.

- [27] S. K. Park, J. D. Venable, T. Xu, and J. R. Yates. A quantitative analysis software tool for mass spectrometry-based proteomics. *Nature Methods*, 5(4), 319-322, 2008.
- [28] S. Dasari, P. A. Wilmarth, A. P. Reddy, L. J. G. Robertson, S. R. Nagalla, and L. L. David. Quantification of Isotopically Overlapping Deamidated and O-18-Labeled Peptides Using Isotopic Envelope Mixture Modeling. *Journal of Proteome Research*, 8(3):1263-1270, 2009.
- [29] T. B. Coplen. Atomic weights of the elements 1999. *Journal of Physical and Chemical Reference Data*, 30(3): 701-712, 2001.
- [30] A. L. Rockwood, S. L. Vanorden, and R. D. Smith. Rapid Calculation of Isotope Distributions. *Analytical Chemistry*, 67(15): 2699-2704, 1995.
- [31] A. L. Rockwood and P. Haimi. Efficient calculation of accurate masses of isotopic peaks. *Journal of the American Society for Mass Spectrometry*, 17(3): 415-419, 2006.
- [32] R. K. Snider. Efficient calculation of exact mass isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 18(8): 1511-1515, 2007.
- [33] M. Wehofsky and R. Hoffmann. Automated deconvolution and deisotoping of electrospray mass spectra. *Journal of Mass Spectrometry*, 37(2): 223-229, 2002.
- [34] J. Fernandez-de-Cossio, L. J. Gonzalez, Y. Satomi, L. Betancourt, Y. Ramos, V. Huerta, V. Besada, G. Padron, N. Minamino, and T. Takao. Automated interpretation of mass spectra of complex mixtures by matching of isotope peak distributions. *Rapid Communications in Mass Spectrometry*, 18(20): 2465-2472, 2004.

- [35] X. J. Li, E. C. Yi, C. J. Kemp, H. Zhang, and R. Aebersold. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Molecular & Cellular Proteomics*, 4(9): 1328-1340, 2005.
- [36] P. Du and R. H. Angeletti. Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Analytical Chemistry*, 78(10): 3385-3392, 2006.
- [37] L. Chen, S. K. Sze, and H. Yang. Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. *Analytical Chemistry*, 78(14), 5006-5018, 2006.
- [38] K. Park, J. Y. Yoon, S. Lee, E. Paek, H. Park, H. J. Jung, and S. W. Lee. Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. *Analytical Chemistry*, 80(19): 7294-7303, 2008.
- [39] J. Y. Yoon, K. Y. Lim, S. Lee, K. Park, E. Paek, U. B. Kang, J. Yeom, and C. Lee. Improved quantitative analysis of mass spectrometry using quadratic equations. *Journal of Proteome Research*, 9(5): 2775-2785, 2010.
- [40] J. Y. Yoon, J. Yeom, H. Lee, K. Kim, S. Na, K. Park, E. Paek, and C. Lee. High-throughput peptide quantification using mTRAQ reagent triplex. *BMC Bioinformatics*, 12(Suppl 1): S46, 2011.
- [41] D. Valkenburg, I. Jansen, and T. Burzykowski. A model-based method for the prediction of the isotopic distribution of peptides. *Journal of the American Society for Mass Spectrometry*, 19(5): 703-712, 2008.
- [42] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J.

- Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic acids research*, 34(suppl 1), D187-D191, 2006.
- [43] J. Klimek, J. S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P. R. Gafken, J. E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossola, J. K. Eng, R. Aebersold, and D. B. Martin. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *The Journal of Proteome Research*, 7(01), 96-103, 2007.
- [44] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry*, 74(20), 5383-5392, 2002.
- [45] A. R. Shah, N. Jaitly, N. Zuljevic, M. E. Monroe, A. Liyu, A. D. Polpitiya, and I. Gorton. An architecture for real time data acquisition and online signal processing for high throughput tandem mass spectrometry. *Fifth IEEE International Conference on e-Sceince*, 88-93, 2009.
- [46] U. B. Kang, J. Yeom, H. J. Kim, H. Kim, and C. Lee. Expression profiling of more than 3500 proteins of MSS-type colorectal cancer by stable isotope labeling and mass spectrometry. *Journal of proteomics*, 75(10), 3050-3062, 2012.
- [47] K. R. Uhlmann, S. Gibb, S. Kalkhof, U. Arroyo-Abad, C. Schulz, B. Hoffmann, and R. Feltens. Species determination of *Culicoides* biting midges via peptide profiling using matrix-assisted laser desorption ionization mass spectrometry. *Parasites & vectors*, 7(1), 1-18, 2014.

- [48] J. Yeom, M. J. Kang, D. Shin, H. K. Song, C. Lee, and J. E. Lee. mTRAQ-based quantitative analysis combined with peptide fractionation based on cysteinyl peptide enrichment. *Analytical biochemistry*, 477, 41-49. 2015.
- [49] I. H. Madar, S. Back, D. G. Mun, H. Kim, J. H. Jung, K. P. Kim, and S. W. Lee. Reduction of Ambiguity in Phosphorylation-site Localization in Large-scale Phosphopeptide Profiling by Data Filter using Unique Mass Class Information. *Bulletin of the Korean Chemical Society*, 35(3), 845-850, 2014.
- [50] J. W. Kim, S. Lee, K. Park, S. Na, E. Paek, H. S. Park, and H. Y. Kim. Monoisotopic Mass Determination Algorithm for Selenocysteine-Containing Polypeptides from Mass Spectrometric Data Based on Theoretical Modeling of Isotopic Peak Intensity Ratios. *Journal of proteome research*, 11(9), 4488-4498, 2012.
- [51] M. Niu, X. Mao, W. Ying, W. Qin, Y. Zhang, and X. Qian. Determination of monoisotopic masses of chimera spectra from high-resolution mass spectrometric data by use of isotopic peak intensity ratio modeling. *Rapid Communications in Mass Spectrometry*, 26(16), 1875-1886, 2012.
- [52] G. W. Slys, E. S. Baker, A. R. Shah, N. Jaitly, G. A. Anderson, and R. D. Smith. The DeconTools framework: an application programming interface enabling flexibility in accurate mass and time tag workflows for proteomics and metabolomics. *In Proc 58th ASMS Conf Mass Spectrom Allied Topics*, 2010.
- [53] C. Y. Hwang, K. Kim, J. Y. Choi, Y. J. Bahn, S. M. Lee, Y. K. Kim, C. Lee, and K. S. Kwon. Quantitative proteome analysis of age-related changes in mouse gastrocnemius muscle using mTRAQ. *Proteomics*, 14(1), 121-132, 2014.

초 록

질량 분석법은 펩타이드 서열을 알아내기 위한 가장 확실하고 강력한 분석 기법 중 하나이다. 막대한 양의 질량 분석 데이터를 수작업으로 분석하는 것은 비현실적이기 때문에 자동화된 질량 분석 데이터의 처리 방법 개발이 필수적이다.

본 논문에서는 질량 분석 데이터의 고속 처리에 대하여 연구한다. 첫째, 펩타이드의 동위 원소 분포에 대한 새로운 수학적 모델을 제시하고, 이를 이용하여 동위 원소 집단 및 단동위 원소 질량을 결정하는 알고리즘을 제안한다. 본 모델에서는 동위 원소 분포에서 인접한 두 피크 세기의 비율과, 인접한 세 피크 세기의 비율 곱의 두 가지의 비율을 사용한다. 이 비율들이 펩타이드 질량에 대한 간단한 함수로 근사될 수 있음을 보이고, 이를 이용한 자동화된 알고리즘을 제시한다. 본 논문에서 제안한 방법은 잘 알려진 THRASH 기반의 프로그램들과 비교하여 분석되었다. 우리의 방법은 THRASH보다 많은 알려진 펩타이드의 질량을 찾았으며, 특히 동위 원소 분포가 에버리진(*averagine*)의 분포로부터 유의미하게 벗어난 펩타이드에 대하여 좋은 결과를 보여주었다. 또 다른 장점은 처리 속도로, 최소 제공법에 기반한 THRASH보다 월등히 빠른 속도를 보여주었다.

둘째, 안정 동위 원소 라벨링 중 2중 mTRAQ 라벨링 실험에서의 중첩된 동위 원소 집단에 대한 새로운 수학적 모델을 제시하고, 이로부터 펩타이드의 정량 분석을 위한 알고리즘을 제안한다. 새로운

알고리즘은 XPRESS를 대신하여 Trans-Proteomic Pipeline 공정에 쉽게 적용이 가능하다. mTRAQ 라벨링된 펩타이드에 대하여 XPRESS보다 정확한 정량비와 더 나은 표준 편차를 보여주었으며, 특히 리신을 포함하지 않은 펩타이드에 대하여 펩타이드의 질량이 증가할수록 큰 차이를 보여주었다.

마지막으로, 3중 mTRAQ 실험에서의 정량 분석을 위한 새로운 알고리즘을 제시한다. 이는 앞선 2중 mTRAQ 중첩 모델의 확장이다. 또한, 펩타이드의 용리 구간 결정을 위한 자동화된 방법을 제시한다. 일부 펩타이드는 비슷한 질량과 용리 시간으로 인하여 용리 구간이 겹칠 수 있다. 이를 정확하게 알아내어 분리하는 것은 정확한 정량 분석을 위하여 필수적이다. 본 알고리즘은 표준 단백질 혼합물 실험을 이용하여 검증되었다.

요약어 : 질량 분석법, 고속 처리, 단동위 원소 질량, 동위 원소 집단, 정량 분석, mTRAQ 라벨링

학번 : 2004-21571